(12) UK Patent Application (19) GB (11) 2607111 (13) A

(43) Date of A Publication 30.11.2022

(21) Application No: 2107733.4

(22) Date of Filing: 28.05.2021

(71) Applicant(s):
Imperial College Innovations Ltd
Level 1 Faculty Building,
Imperial College Exhibition Road, London,
Greater London, SW7 2AZ, United Kingdom

(72) Inventor(s):
Mikolaj Jankowski
Tze-Yan Tung
David Burth Kurka
Deniz Gunduz

(74) Agent and/or Address for Service:
HGF Limited
1 City Walk, LEEDS, LS11 9DX, United Kingdom

(51) INT CL:
G06N 3/04 (2006.01) ***G06N 3/08*** (2006.01)
H04N 19/124 (2014.01)

(56) Documents Cited:
WO 2020/035684 A1 US 20200304802 A1
IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, vol 1, 2020, LEINONEN MARKUS ET AL, "Low-Complexity Vector Quantized Compressed Sensing via Deep Neural Networks", pages 1278-1294, DOI:10.1109/OJCOMS.2020.3020131
ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 2018, JUN HAN ET AL, "Deep Probabilistic Video Compression", available from https://arxiv.org/pdf/1810.02845v1.pdf

(58) Field of Search:
INT CL **G06N, H04N**

(54) Title of the Invention: **Communication system for conveying data from an information source across a communication channel**
Abstract Title: **Encoder and decoder neural networks for use in a communication system**

(57) Communication system 100 conveys data from transmitter 110 across communication channel 120 using joint source and channel coding. Transmitter 110 has encoder neural network 115 for encoding input data direct to values in a signal space for modulating a carrier signal for transmission of a transformed version of the input data across a communication channel. Communication system 100 also comprises receiver 130 having decoder neural network 135 for decoding a noise-affected version of vector signal values transmitted over the noisy communication channel direct to a reconstructed representation of the input data provided from the information source. The transmitter further includes soft quantization module 116 and hard quantization module 117 for assigning the output of the encoder to signal values in a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the transmitter over the communication channel. Training the encoder and decoder neural networks uses an appropriate optimization algorithm, e.g. gradient descent, operating on an objective function characterising a reconstruction error between the input-output pairs of training data, consisting of information source samples, which are passed to the encoder neural network, and the corresponding reconstructions at the decoder output layer of the decoder.

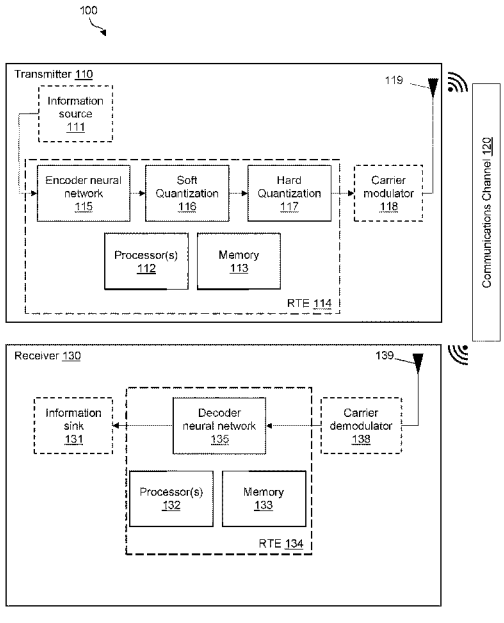

Figure 1

GB 2607111 A

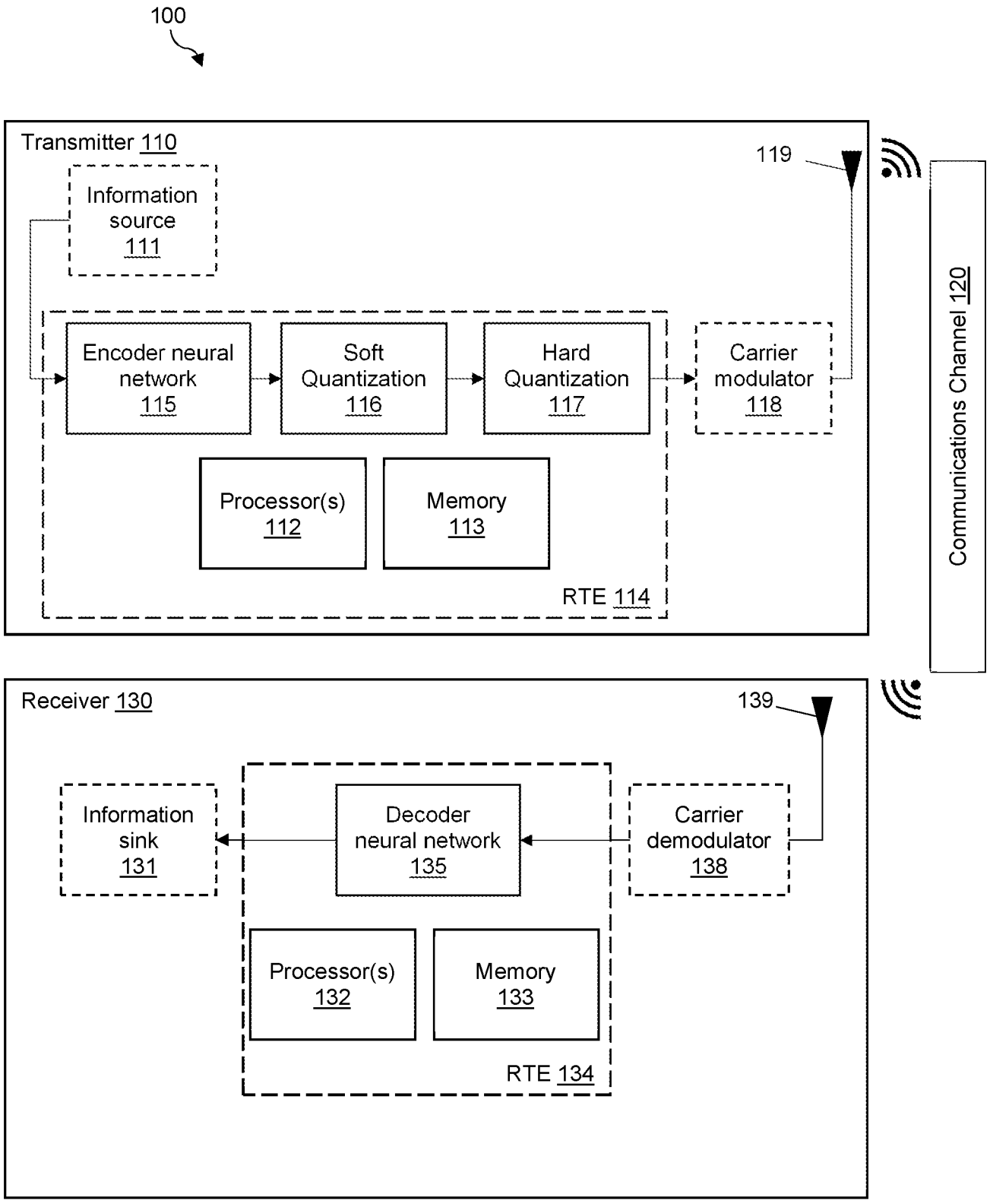
100
Transmitter 110
119
Information source 111
Encoder neural network 115
Soft Quantization 116
Hard Quantization 117
Carrier modulator 118
Processor(s) 112
Memory 113
RTE 114
Communications Channel 120
Receiver 130
139
Information sink 131
Decoder neural network 135
Carrier demodulator 138
Processor(s) 132
Memory 133
RTE 134

Figure 1

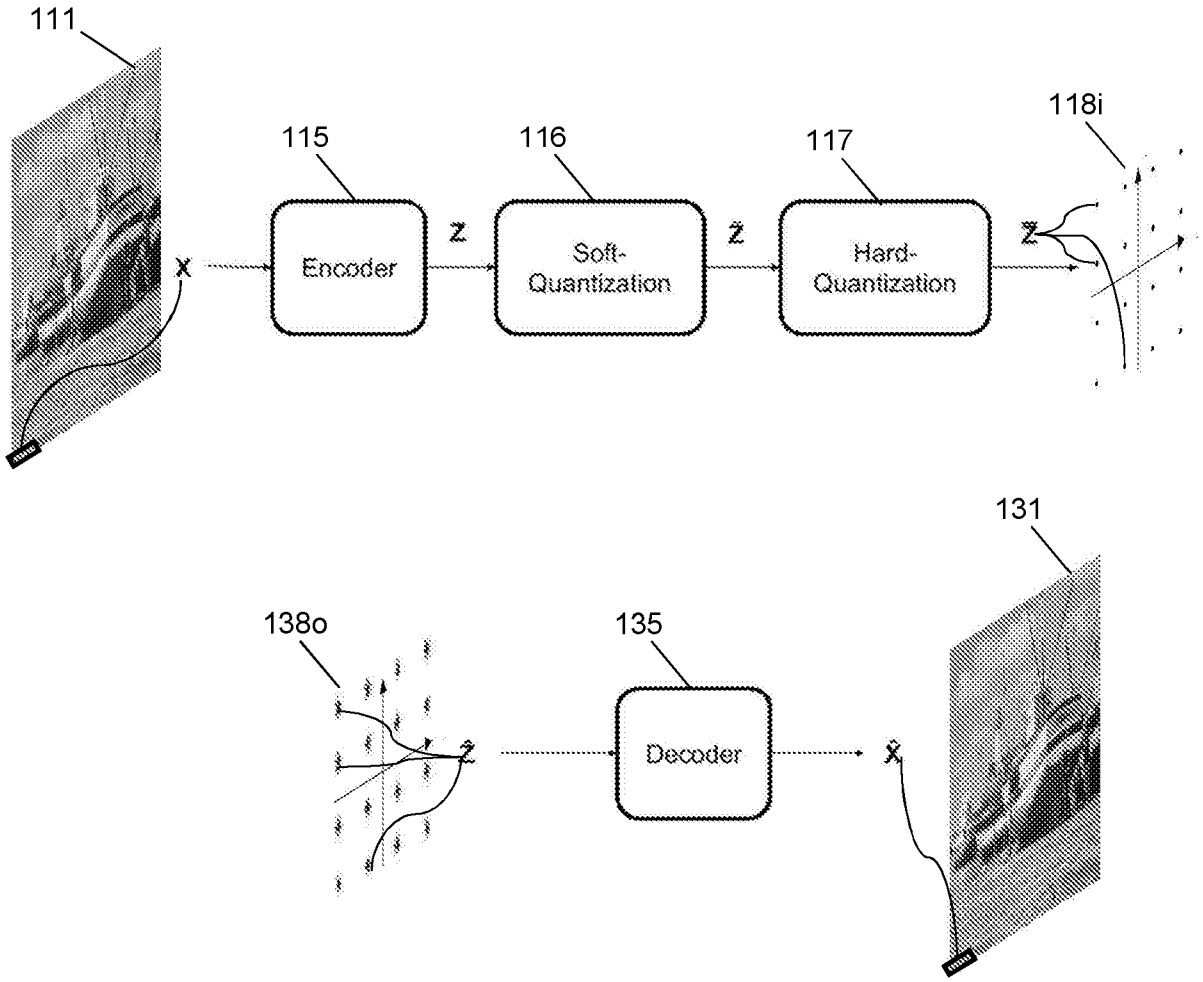
111
115
116
117
118i
Encoder
Soft-
Quantization
Hard-
Quantization
131
138o
135
Decoder

Figure 2

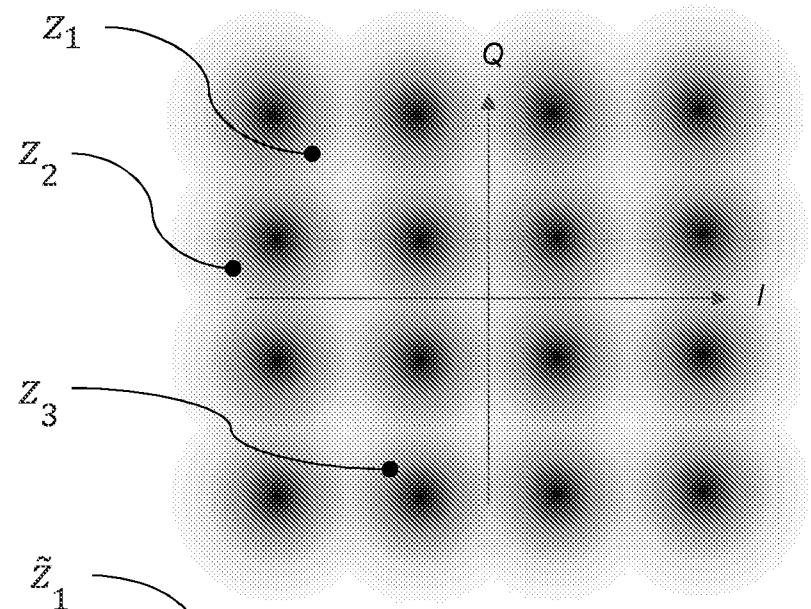

Figure 3A

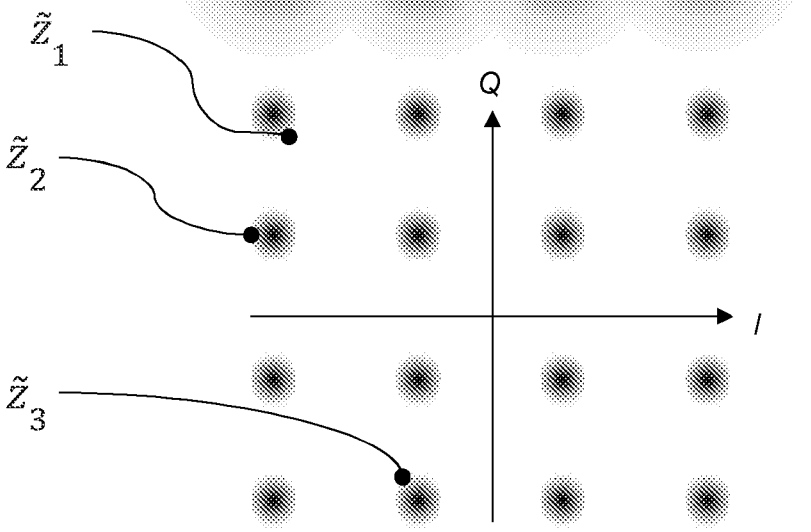

Figure 3B

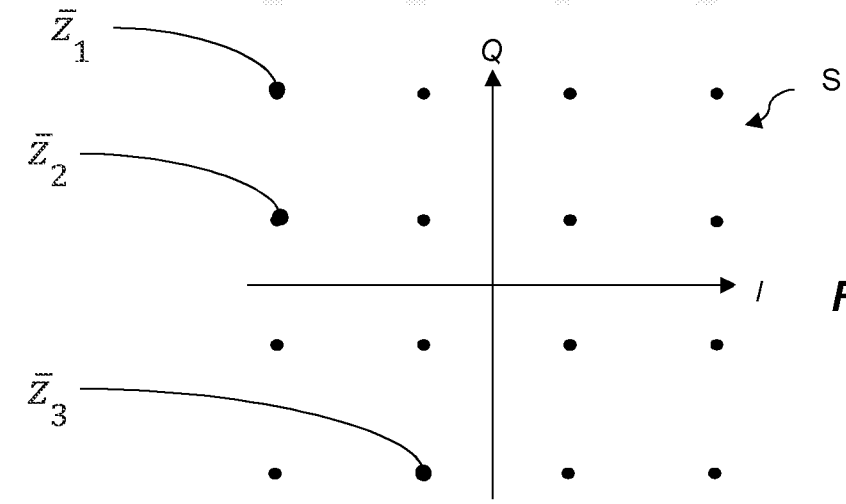

Figure 3C

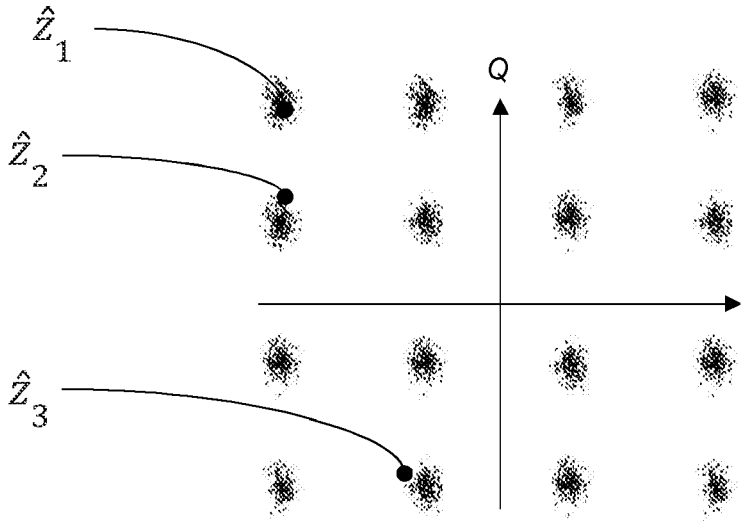

Figure 3D

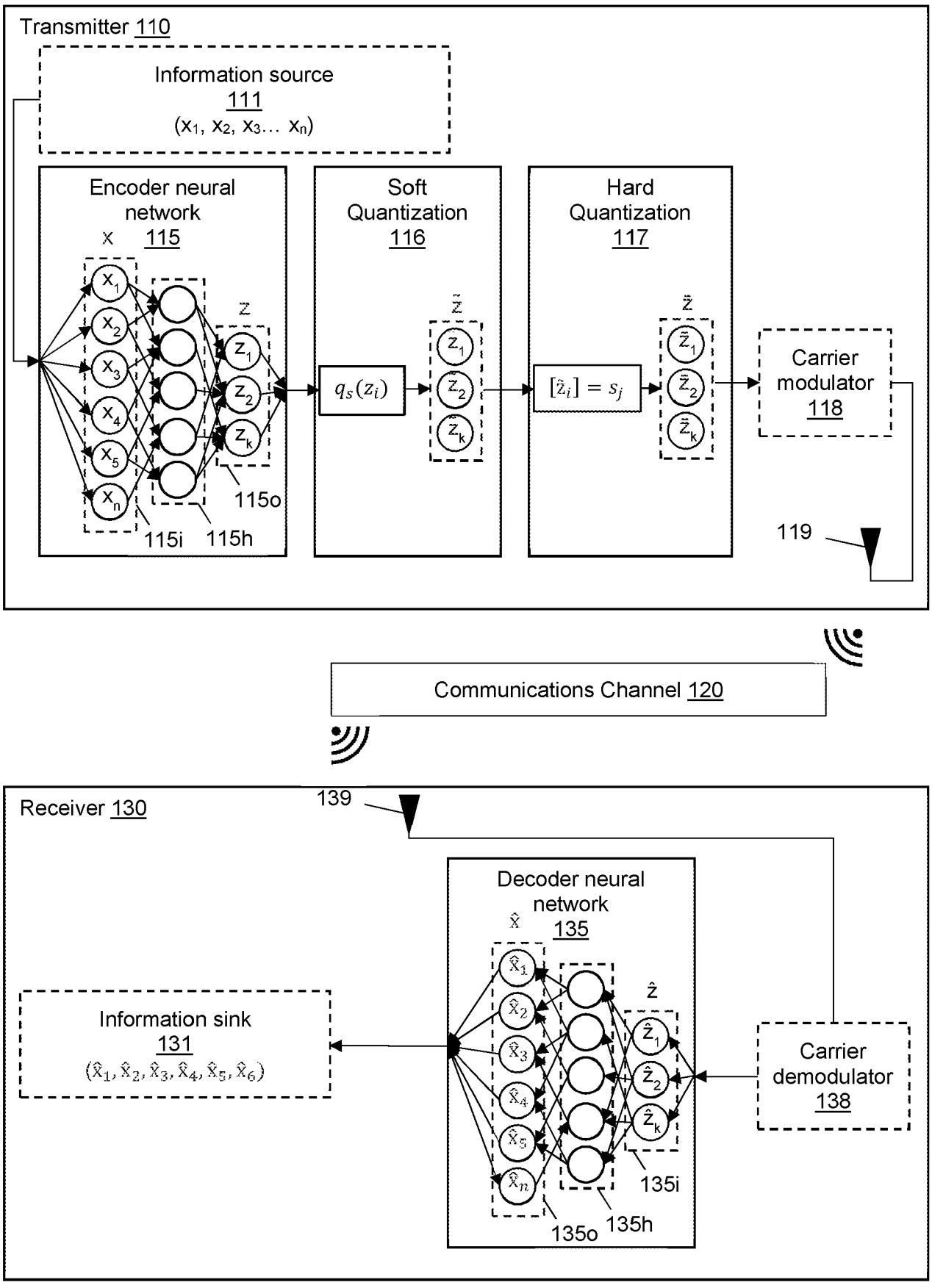
Transmitter 110
Information source
111
(x1, x2, x3… xn)
Encoder neural network
115
Soft Quantization
116
Hard Quantization
117
Carrier modulator
118
119
115o
115i
115h
Communications Channel 120
Receiver 130
139
Decoder neural network
135
Information sink
131
Carrier demodulator
138
135i
135o
135h

Figure 4

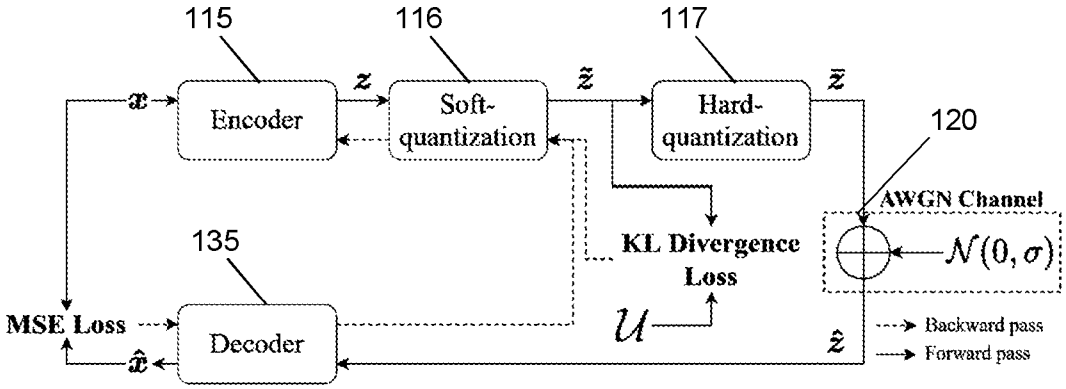

Figure 5

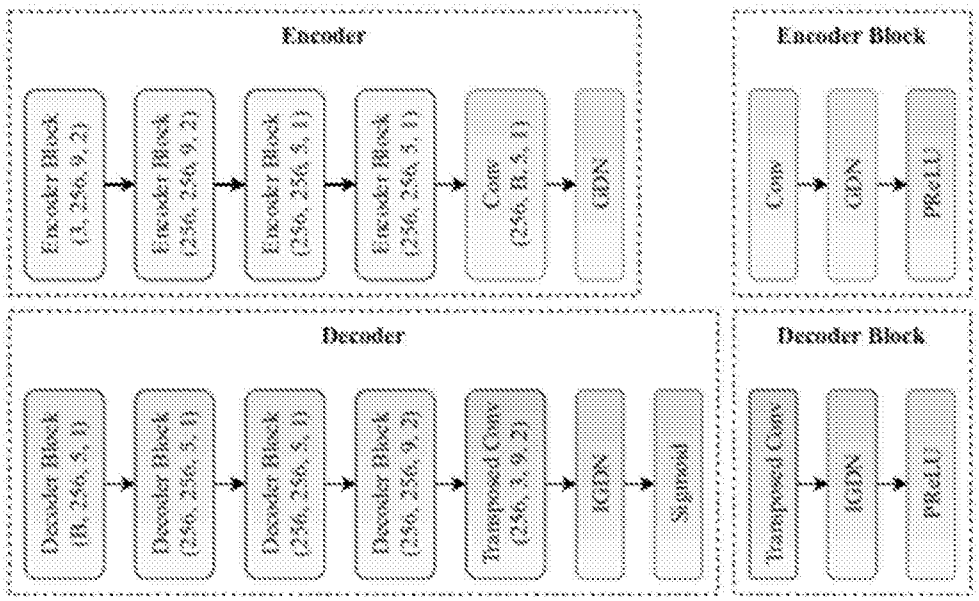

Figure 6

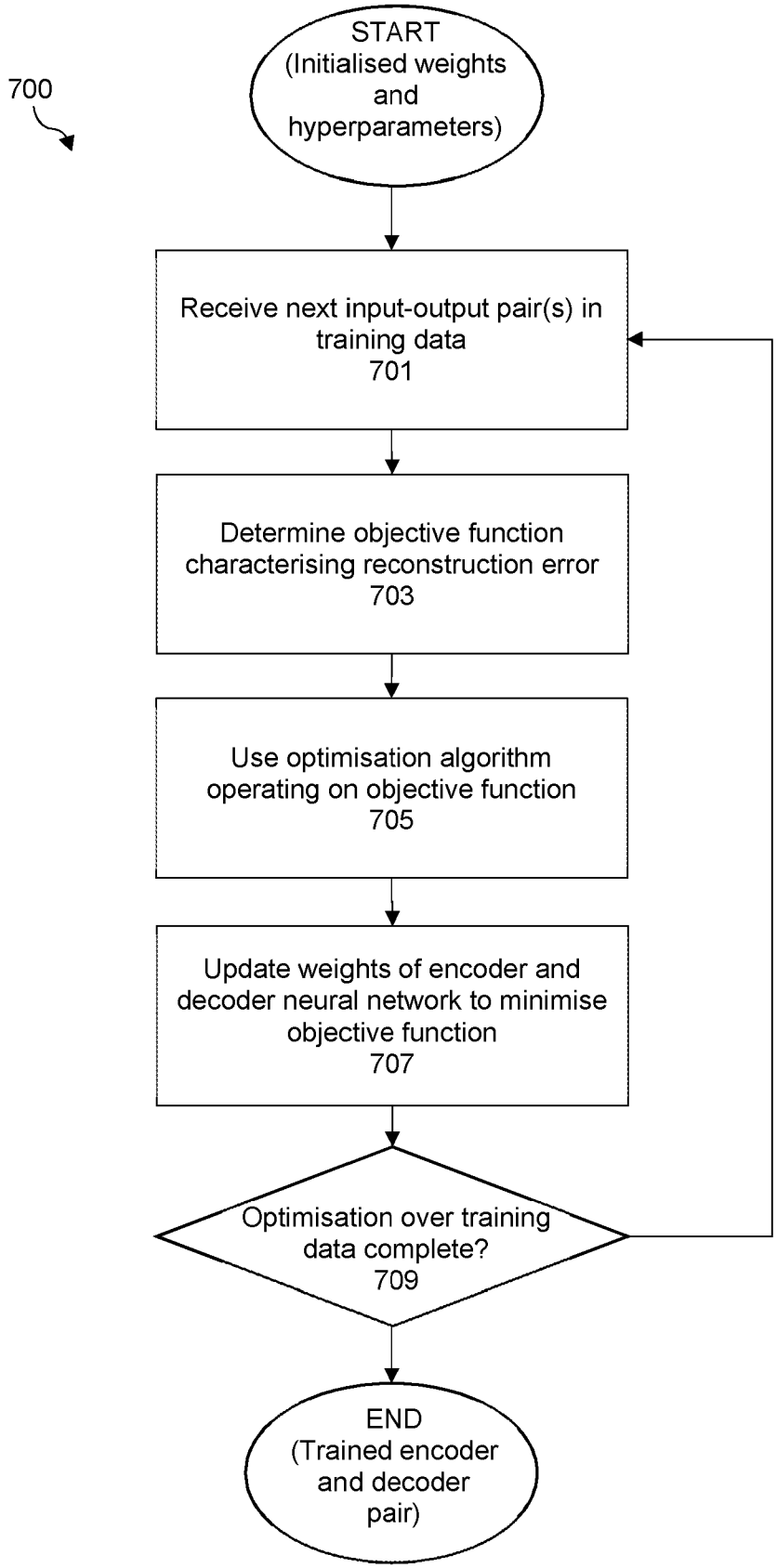
700
START
(Initialised weights and hyperparameters)
Receive next input-output pair(s) in training data
701
Determine objective function characterising reconstruction error
703
Use optimisation algorithm operating on objective function
705
Update weights of encoder and decoder neural network to minimise objective function
707
Optimisation over training data complete?
709
END
(Trained encoder and decoder pair)

Figure 7

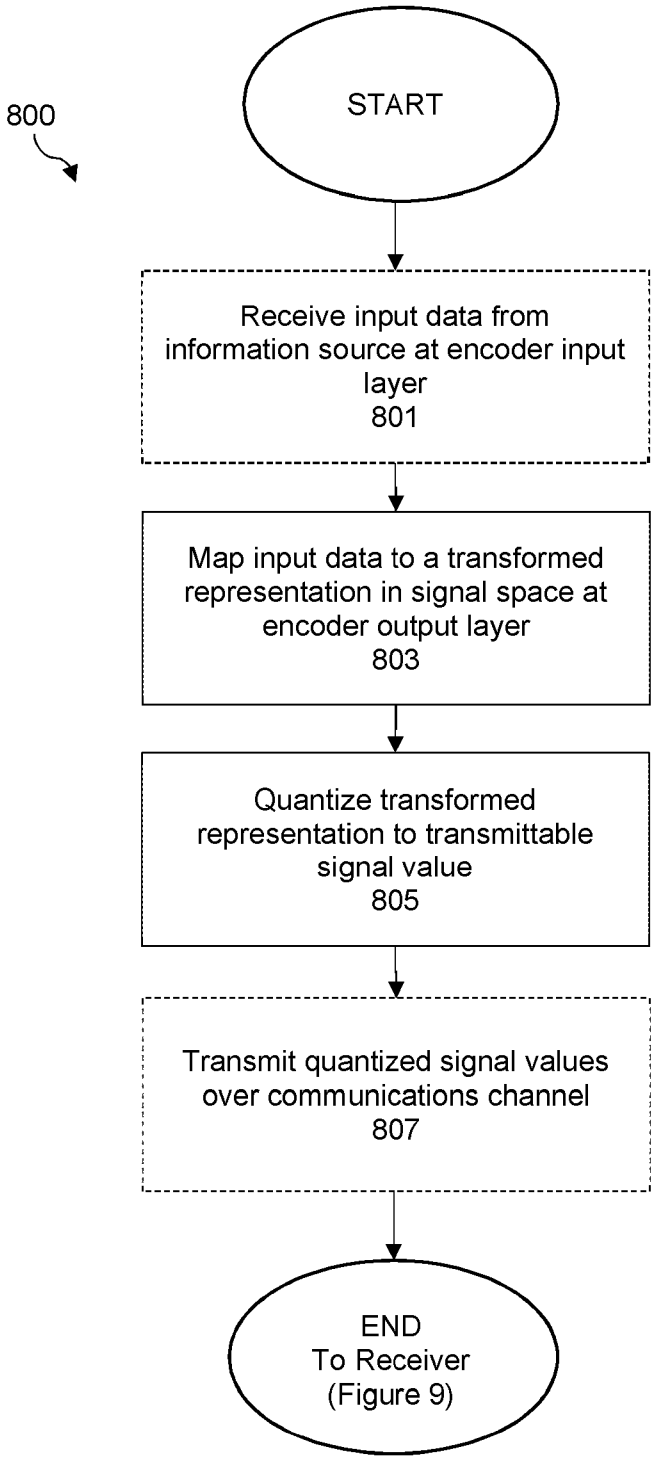
800
START
Receive input data from information source at encoder input layer
801
Map input data to a transformed representation in signal space at encoder output layer
803
Quantize transformed representation to transmittable signal value
805
Transmit quantized signal values over communications channel
807
END
To Receiver
(Figure 9)

Figure 8

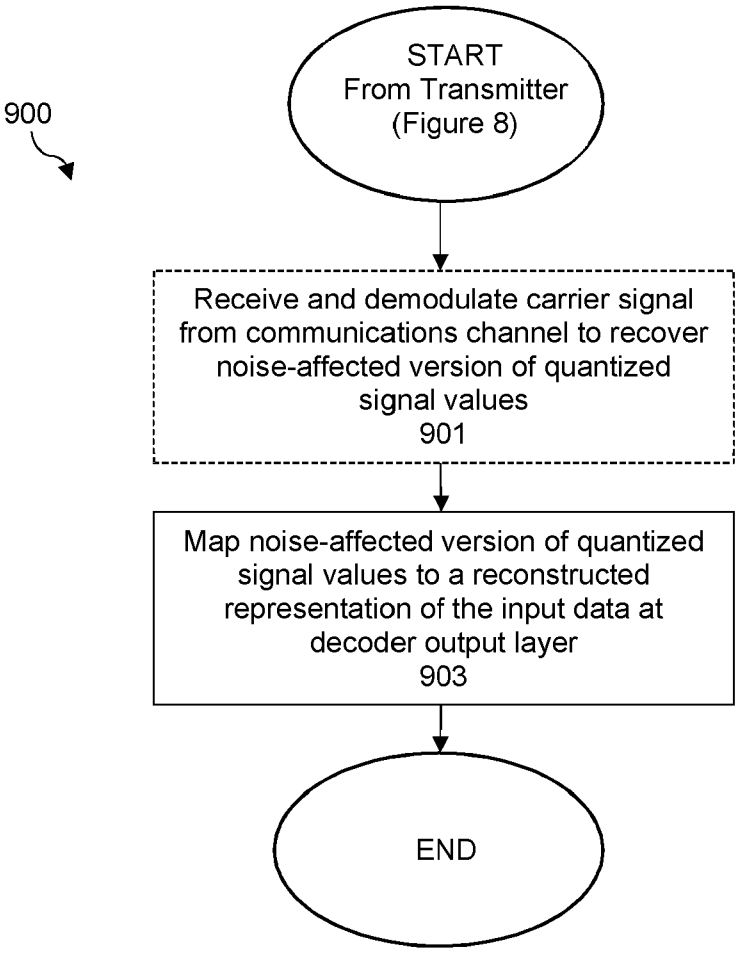
900
START
From Transmitter
(Figure 8)
Receive and demodulate carrier signal from communications channel to recover noise-affected version of quantized signal values
901
Map noise-affected version of quantized signal values to a reconstructed representation of the input data at decoder output layer
903
END

Figure 9

COMMUNICATION SYSTEM, TRANSMITTER, AND RECEIVER FOR CONVEYING DATA FROM AN INFORMATION SOURCE ACROSS A COMMUNICATION CHANNEL USING JOINT SOURCE AND CHANNEL CODING, AND METHOD OF TRAINING AN ENCODER NEURAL NETWORK AND DECODER NEURAL NETWORK FOR USE IN A COMMUNICATION SYSTEM

[0001] This present application relates to a communication system comprising a transmitter and receiver for conveying data from an information source across a communication channel. The transmitter and receiver include a pair of encoder and decoder neural networks for joint source and channel coding which have been jointly trained to reduce reconstruction errors.

BACKGROUND

[0002] An aim of a data communication system is to efficiently and reliably send data from an information source over a communication channel from a transmitter at as high a rate as possible with as few errors as achievable in view of the channel noise, to enable a faithful representation of the original information source to be recovered at a receiver.

[0003] Most digital communication systems today include a source encoder and separate channel encoder at a transmitter and a source decoder and separate channel decoder at a receiver.

[0004] In digital communication systems, to transmit data from the information source over a communication channel, the symbols of source data are first digitally compressed into bits by the source encoder. The goal in source coding is to encode the sequence of source symbols into a coded representation of data elements to reduce the redundancy in the original sequence of source symbols. In lossless compression one has to remove redundancy such that the original information source can still be reconstructed as the original version from the coded representation, while lossy compression allows a certain amount of degradation in the reconstructed version under some specified distortion measure, for example squared error. JPEG for images, or H264/MPEG for videos are examples of lossy source compression standards widely used in practice. Compressing the information source using a source encoder before transmission means that fewer resources are required for that transmission.

[0005] Once the data from the information source has been encoded to compress it down in size, to transfer this representation over a communication channel, the output of the source encoder is then provided to a channel encoder. The goal of the channel encoder is to encode the compressed data representation in a structured way using a suitable Error Correction Code (ECC) by *adding* redundancy such that even if some of these bits are distorted or lost due to noise over the channel, the receiver can still recover the original sequence of bits reliably. The amount of redundancy that is added depends on the statistical properties of the underlying

communication channel and a target Bit Error Rate (BER). Generally, such channel coding schemes using Forward Error Correction (FEC) provide for a faithful recovery of the transmitted data elements (such as a compressed data source) where the noise on the channel leads to a quality of signal reception below a maximum BER. However, due to the cliff effect, as channel noise increases, BER will increase drastically and, when the channel noise is too high and a maximum BER is breached, the signal transmission will drop out completely, meaning the transmitted data cannot be recovered. There are many different channel coding techniques in practice that provide various complexity and performance trade-offs. Turbo codes and Low-density parity-check (LDPC) codes are examples of ECCs that are commonly used in modern communication systems such as WiMAX and fourth generation Long-Term Evolution (LTE) mobile communications.

[0006] The coded bits at the output of the channel encoder are transmitted over the channel using a modulator. The modulator converts the bits into signals that can be transmitted over the communication medium. For example, in wireless systems using Quadrature Modulation of two out-of-phase amplitude modulated carrier signals, the transmitted waveform is specified by its In-Phase (I) and Quadrature (Q) components, and a modulator typically has a discrete set of pre-specified I and Q values, called a constellation, and each group of coded information bits are mapped to a single point in this constellation. Example modulation schemes include phase shift keying (PSK) and quadrature amplitude modulation (QAM).

[0007] The receiver receives and demodulates (for example, by coherent demodulation) a sequence of noisy symbols, where the noise has been added by the communication channel. These noisy demodulated symbols are then mapped to sequences of data elements by a channel decoder. The decoded data elements are then passed to the source decoder, which decodes these data elements to try to reconstruct a representation of the original input source symbols to reconstruct the information source.

[0008] Naturally, the source encoder and decoder are designed jointly, as are the channel encoder and decoder, but the source encoder/decoder and channel encoder/decoder are designed and operate separately to perform very different functions.

[0009] The main advantage of separate source and channel coding is the modularity it provides. This means that the same channel encoder and decoder can be used in conjunction with any source encoder and decoder. That is, as long as the source encoder outputs data elements that can be encoded by the channel encoder, it does not matter if these bits come from an image compressor or a video encoder. Thus, a channel encoder can encode data elements for transmission over a channel irrespective of the data elements or the information source from which they have been derived.

[0010] Similarly, the source encoder and decoder can be operated in conjunction with any channel encoder and decoder to transmit the encoded source symbols over a communication channel. Thus, a source encoder can encode data elements for subsequent coding by the channel encoder independently of which channel encoder is used.

[0011] Thus, the current communication protocols have been developed over generations, and their wide adoption has led to the standardization of some aspects of it. For instance, the task of creating information symbols to be transmitted over a complex channel have typically followed the steps of discretizing (or quantizing) data from an information source into a finite set of symbols through source coding, which are then encoded against channel noise and mapped to symbols from a fixed discrete set of complex representations, also called as a *constellation set of a constellation diagram,* such as the quadrature amplitude modulation (QAM) or binary phase-shift keying (BPSK), through channel coding.

[0012] These choices, based on results and assumptions that go back to the early times of information theory and derived through expert knowledge and handcrafted optimization techniques, have long been adopted by the communication standards. For example, the 5G New Radio standard allows the devices to choose from 16-QAM, 64-QAM and 256-QAM, whereas the most recent Wi-Fi 6 (802.11ax) standard employs 1024-QAM. The constellation size directly impacts the transmission rate, and hence, the constellation sizes are much larger in wired connections. For example, Ethernet devices may use 1024-QAM or 4096-QAM, while in the ADSL technology the constellation size may go up to 32768-QAM. Given their prevalence in standards, current communication hardware is typically hard-wired to work only with these kind of symbols. Such hardware standardisation also makes circuits simpler, allowing mass production and saving costs, leading to their widespread adoption to deliver reliable, high-throughput data communications.

[0013] It is in the above context that the present disclosure has been devised.

BRIEF SUMMARY OF THE DISCLOSURE

[0014] Recently, new alternatives for the design of wireless communication systems have been proposed. For example, machine learning (ML) approaches to encoding an information source to channel symbols for transmission across a communication channel have been proposed. By using ML in this way, new encoding schemes can be discovered and freely produced that optimise the efficient transmission of an information source across a noisy communication channel, without being limited to existing source or channel coding paradigms, often outperforming these legacy handcrafted approaches.

[0015] As such new communication systems are being invented, they often lead to a change in encoding paradigm and that can lose compatibility with existing hardware and standards. This hinders the proliferation of such strategies where they find encoding schemes that are incompatible with existing communication hardware and standards. The present disclosure provides an approach to providing communication systems that can benefit from the discovery of new, ML-optimised communication strategies, while maintaining a common interface with a wide variety of existing devices, that can be deployed on existing communication hardware.

[0016] Thus, viewed from one aspect, the present disclosure provides a communication system for conveying data from an information source across a communication channel using joint source and channel coding, comprising a transmitter and a receiver.

[0017] In one aspect, the transmitter comprises an encoder neural network having: an encoder input layer having input nodes for receiving input data from an information source to be transmitted; an encoder output layer having output nodes for outputting a vector of values in a signal space for modulating a carrier signal for transmission of the input data across a communication channel; and one or more hidden layers having connecting nodes connecting the encoder input layer to the encoder output layer, the connecting nodes having weights to map the input data received at the encoder input layer to provide at the encoder output layer a transformed representation of the input data. The transmitter further comprises a soft quantization module configured to receive the unconstrained output signal space values from the encoder output layer and to operate a differentiable function over the signal space that transforms the unconstrained output signal space values to a smoothed approximation of a signal value in a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the transmitter over the communication channel. The transmitter also further comprises a hard quantization module configured to transform the approximate signal values output from the soft quantization module exactly to a signal value assigned to the symbol in the predetermined finite set S of the alphabet, the hard quantized signal values representing a transformed version of the input data and being passed as a vector from the hard quantization module to the transmitter to modulate a carrier signal for transmission over the communication channel.

[0018] In one aspect, the receiver comprises a decoder neural network having: a decoder input layer having nodes corresponding to a channel output vector received at the receiver receiving the signal, the channel output vector corresponding to a noise-affected version of the vector of hard quantized signal values transmitted over the noisy communication channel; a decoder output layer having output nodes for outputting a reconstructed representation of the input data provided from the information source to the encoder input layer of the encoder; and one or more hidden layers having connecting nodes connecting the decoder input layer to the

decoder output layer, the connecting nodes having weights to map the noise-affected version of the vector of hard quantized signal values received at the decoder input layer to provide at the decoder output layer a reconstructed representation of the input data.

[0019] In the communication system, the connecting node weights of the hidden layers of the encoder neural network and decoder neural network have been trained together to seek to minimise an objective function characterising a reconstruction error between the input-output pairs of training data, consisting of information source samples, which are passed to the encoder neural network, and the corresponding reconstructions at the decoder output layer of the decoder, using an appropriate optimization algorithm operating on the objective function.

[0020] Viewed from another aspect, the present disclosure provides a method for conveying data from an information source across a communication channel using joint source and channel coding. The method comprises, at a transmitter, receiving input data from an information source to be processed at nodes of an encoder input layer of an encoder neural network, and mapping the input data received at the encoder input layer to provide at the encoder output layer a transformed representation of the input data, the encoder output layer having output nodes for outputting a vector of values in a signal space for modulating a carrier signal for transmission of a transformed version of the input data across a communication channel. The mapping is performed by one or more hidden layers of the encoder neural network having connecting nodes connecting the encoder input layer to an encoder output layer, the connecting nodes having weights to map the input data to a transformed representation of the input data in a vector of values in the signal space. The method further comprises operating, by a soft quantization module of the transmitter, receiving the unconstrained output signal space values from the encoder output layer and operating a differentiable function over the signal space that transforms the unconstrained output signal space values to a smoothed approximation of a signal value in a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the transmitter over the communication channel. The method further comprises, by a hard quantization module of the transmitter, transforming the approximate signal values output from the soft quantization module exactly to a signal value assigned to the symbol in the predetermined finite set S of the alphabet. The hard quantized signal values represent the transformed version of the input data and are passed as a vector from the hard quantization module to the transmitter to modulate a carrier signal for transmission over the communication channel.

[0021] Viewed from another aspect, the present disclosure provides a method for conveying data from an information source across a communication channel using joint source and channel coding. The method comprises, at a receiver, receiving the signal transmitted from the

transmitter and passing the received and demodulated signal as a channel output vector to a decoder input layer of a decoder neural network, the decoder input layer having nodes corresponding to the channel output vector, the channel output vector corresponding to a noise-affected version of the vector of hard quantized signal values transmitted over the noisy communication channel by the transmitter. The method further comprises, at output nodes of a decoder output layer of the decoder neural network, outputting a reconstructed representation of the input data provided from the information source to the encoder input layer of the encoder. The reconstruction of the input data is achieved by one or more hidden layers having connecting nodes connecting the decoder input layer to the decoder output layer, the connecting nodes having weights to map the noise-affected version of the vector of hard quantized signal values received at the decoder input layer to provide at the decoder output layer a reconstructed representation of the input data.

[0022] In the methods, the connecting node weights of the hidden layers of the encoder neural network and decoder neural network have been trained together to seek to minimise an objective function characterising a reconstruction error between the input-output pairs of training data, consisting of information source samples, which are passed to the encoder neural network, and the corresponding reconstructions at the decoder output layer of the decoder, using an appropriate optimization algorithm operating on the objective function.

[0023] In this regard, viewed from another aspect, the present disclosure provides a method of training an encoder neural network and decoder neural network for use in a communication system as described above, for conveying data from an information source across a communication channel using joint source channel coding, the method comprising: for input-output pairs of a set of training data from the information source passed to the encoder neural network, determining an objective function characterising a reconstruction error between the input-output pairs of training data, consisting of information source samples, which are passed to the encoder neural network, and the corresponding reconstructions at the decoder output layer of the decoder, using an appropriate optimization algorithm operating on the objective function, updating the connecting node weights of the hidden layers of the encoder neural network and decoder neural network to seek to minimise the objective function.

[0024] In accordance with examples of the present disclosure, ML-based communication systems for conveying data from an information source across a communication channel are provided that can be incorporated directly into existing conventional and commercial communication hardware. This is achieved by using a system that provides an ML-optimised joint source and channel coding scheme that can transform data from an information source (e.g., image or video) into a discrete set of signal values in a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the

transmitter over the communication channel utilising existing transceiver hardware and communication protocols. The ML-optimised decoder can then map the received noisy channel output back to a reconstruction of the information source at the receiver side. As the ML-designed encoding schemes of the present disclosure translates data from an information source directly to channel symbols belonging to a set (or constellation) of standardized communication symbols, this designed joint source and channel coding modules can be incorporated into existing communication hardware that already operates with those symbols.

[0025] This is enabled by the incorporation in the transmission chain of the quantization modules. In particular, the provision of the soft and then hard quantization modules allow in the forward pass of data through the encoder at training time and at runtime the exact signal values in a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the transmitter to be output. At the same time, although the encoded output of the transmitter from the hard quantization module is exact, non-differentiable signal values, the soft quantization module transforms the output of the encoder neural network to a smoothed approximation of signal values in the predetermined finite set S of symbols for transmission. This smoothed approximation allows, in a backward pass of the data through the decoder and encoder neural networks, a differentiable objective function characterising the reconstruction error to be obtained, which is then permits the optimisation of the connecting node weights of the encoder and decoder neural networks to be performed in a training phase, allowing the reconstruction error of the communication system to be minimised.

[0026] The separation of quantization into soft and hard steps enables the end-to-end learning of the communication system's processing blocks using, for example, a stochastic gradient descent optimisation method over a training set, allowing an ML-optimised joint source and channel coding scheme to be learned for existing communication hardware and protocols, despite the output of the transmitter being quantized in a non-differentiable symbol assignment.

[0027] It should be noted that, in accordance with the present disclosure, the communication channel should be understood as any transformation from the channel input space to the channel output space that includes a random transformation due to the channel. This may include additive noise, interference, or other stochastic properties of the channel that will randomly transform the transmitted signal, e.g., fading and multi-path effects in wireless channels. Thus, the reference to the noise-affected version of the of the vector of hard quantized signal values received at the decoder should be understood to indicate that the input to the decoder is a vector of values correlated with the transmitted vector of signal values (which is itself correlated with the input data from the information source), transformed by the communications channel, whether that transformation is 'noise' or another channel transformation.

[0028] In this respect, in accordance with the present disclosure, the communications channel should be understood as encompassing any channel that has, in an input signal space, a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the transmitter over the communication channel (which applies a random transformation to the channel output space). Thus, the communication channel may be one that also includes an existing channel encoder and decoder scheme, in which the signal space of the channel input may be the predetermined finite set symbols of the channel code (which could be bit values) for modulating a signal carrier for providing input signals in the alphabet in the input signal space for the communication channel. Thus, besides random noise applied by the communication channel, the transformation applied by the communication channel may also include an existing channel code. Thus, in these embodiments, the predefined alphabet of symbols for quantization may be a message alphabet for an existing channel code by which the input signals to the given communications channel are modulated. In this case, the hard quantized channel inputs will be mapped into the message alphabet of the corresponding channel code. The noise-affected channel output may correspond to the hard-decoded message of the channel decoder. In this respect, in these embodiments, the encoder and decoder neural networks may learn an optimum mapping of the input information source to inputs of an existing channel code of the communications channel that reduces reconstruction errors at the decoder. Although acting as an outer code in these embodiments, this learned coding is still optimised based on the characteristics of the channel to reduce reconstruction errors, even though the communication channel includes an existing channel code. This is unlike existing modular source codes which are defined independently of the random transformation applied by any channel.

[0029] Further, in accordance with the present disclosure, it should be understood that the dimension of the vector of hard-quantized signal values that are transmitted over the communication channel need not be equal to the dimension of the channel output vector corresponding to the noise-affected version of the vector of hard quantized signal values transmitted over the noisy communication channel. That is, in embodiments the version of the vector of hard quantized signal values transmitted over the noisy communication channel that is input into the decoder input layer may have a dimensionality different to the transmitted vector of hard quantized signal values. This may be due, for example, to characteristics of the receiver, the detection and demodulation process revealing a different signal vector. For example, this may be due to different number of receive elements at the receiver, e.g., multiple antennas, or a different sampling rate of the received signal. The difference in dimensionality of the version of the transmitted signal input to the decoder compared to the quantized signal output at the transmitter may also be due to some processing carried out at the receiver. For example, each detected signal may be provided to the decoder as a vector of likelihoods of

assignment to the different symbol values for the communications channel output. That is, the decoder may receive, for each detected signal value, a vector of soft decisions on assignment to the different symbols, rather than a specific symbol value. In this respect, the noise-affected version of the hard quantized signal values provided to the decoder may represent the received signal information differently to the representation provided to the modulator at the transmitter. Nevertheless, regardless of its dimensionality and representation of the detected signal values, the noise-affected version of the vector of hard quantized signal values transmitted over the noisy communication channel and provided to the decoder is simply one that is correlated with the input data from the information source. The decoder then learns to reconstruct the information source from this noise-affected version of the transmitted signals, regardless of its dimensionality and representation.

[0030] It should also be noted that, in accordance with the present disclosure, the soft and hard quantization steps can be carried out symbol-by-symbol or over blocks of symbols. That is, the input to the soft quantization module may be plural latent vector elements in the signal space, which may be quantized together to provide plural soft quantized outputs. Thus, it should be understood that the quantization provided in accordance with the present disclosure can be scalar quantization or vector quantization.

[0031] In examples of the present disclosure, the hard quantized signal values represent in-phase and quadrature components for modulation of the carrier signal for transmission over the communication channel.

[0032] In examples of the present disclosure, wherein the predefined alphabet is a fixed, predefined constellation of symbols for digitally modulating the carrier signal to encode the input data for transmission over the communication channel.

[0033] In examples of the present disclosure, the predefined alphabet is selected to be one that compatible with the modulator of the transmitter and/or compatible with the demodulator of the receiver.

[0034] In examples of the present disclosure, the alphabet of symbols is mapped onto a constellation diagram for Phase Shift Keying or Quadrature Amplitude Modulation of the carrier signal.

[0035] In examples of the present disclosure, the encoder neural network and decoder neural network have been trained together using a gradient descent algorithm operating on the objective function based on a differential of at least the decoder neural network, the soft quantization function and the encoder neural network.

[0036] In examples of the present disclosure, the encoder neural network and decoder neural network have been trained together using training data in which a model of the communication

channel is used to model the effect of the random channel on the transmitted hard quantized signal values to generate a noise-affected version of the vector of hard quantized signal values in the input-output pairs of training data.

[0037] In examples of the present disclosure, during training the gradient descent algorithm has further operated on the objective function based on a differential of the channel model. In examples of the present disclosure, during training the gradient descent algorithm has operated on the objective function further based on a differentiable representation of the channel model generated using a generative adversarial network. In this way, the encoder and decoder neural networks can be optimised to take into account the noise in the channel through training based on a model of the communication channel not-necessarily explicit or differentiable.

[0038] In examples of the present disclosure, the encoder neural network and decoder neural network have been trained together using training data in which input-output pairs of training data are transmitted over the communication channel by the transmitter, in order to add noise to the transmitted hard quantized signal values to generate the noise-affected version of the vector of hard quantized signal values in the input-output pairs of training data. In this way, the encoder and decoder neural networks can be optimised to take into account the noise in the channel through training based on an empirical data capturing the effects of channel statistics on the transmission.

[0039] In examples of the present disclosure, the symbol alphabet set size S is a parameter of the soft and hard quantization modules, and wherein during training the gradient descent algorithm has further operated over a range of symbol alphabet set sizes S of different cardinality to learn the optimal modulation scheme for the information source. In this way, the optimal transmittable constellation for transmitting the information source across the communication channel can be learned.

[0040] In examples of the present disclosure, the objective function sought to be minimised by the training together of the encoder neural network and decoder neural network additionally characterises a relative entropy between the probability distribution of the symbols of the hard quantized signal values output from the hard quantization module and a uniform distribution across the symbols. In examples of the present disclosure, the Kullback-Liebler divergence between the symbol distribution and the uniform distribution is used in the objective function to characterise the relative entropy for minimisation during training. In this way, the encoder and decoder can be trained to ensure that, in the transmission of the symbols representing data from the information source across the communication channel, an average power constraint on the communication channel is satisfied.

[0041] In examples of the present disclosure, the information source is uncompressed. In examples of the present disclosure, the encoder neural network maps the input data to a

compressed representation of the input data from the information source to a vector of values in the signal space.

[0042] In examples of the present disclosure, the latent variables are produced by the encoder neural network are unconstrained in the signal space.

[0043] It will be appreciated from the foregoing disclosure and the following detailed description of the examples that certain features and implementations described as being optional in relation to any given aspect of the disclosure set out above should be understood by the reader as being disclosed also in combination with the other aspects of the present disclosure, where applicable. Similarly, it will be appreciated that any attendant advantages described in relation to any given aspect of the disclosure set out above should be understood by the reader as being disclosed as advantages of the other aspects of the present disclosure, where applicable. That is, the description of optional features and advantages in relation to a specific aspect of the disclosure above is not limiting, and it should be understood that the disclosures of these optional features and advantages are intended to relate to all aspects of the disclosure in combination, where such combination is applicable.

BRIEF DESCRIPTION OF THE DRAWINGS

[0044] Embodiments of the invention are further described hereinafter with reference to the accompanying drawings, in which:

Figure 1 shows a communication system for conveying data from an information source across a communication channel using joint source and channel coding in accordance with an example of the present disclosure;

Figure 2 shows the encoding, transmission and reconstruction of an image from an information source using a communication system in accordance with an example of the present disclosure;

Figure 3A shows an example vector of three values, z_1, z_2, z_3 each having in-phase and quadrature components in an IQ signal space for modulating a carrier signal for transmission across a communication channel to which a vector of input data from an information source has been mapped by an encoder neural network in accordance with an example of the present disclosure;

Figure 3B shows an example vector of three values $\tilde{z}_1, \tilde{z}_2, \tilde{z}_3$ in the signal space to which the encoded signal values z_1, z_2, z_3 have been soft quantized by a soft quantization module in accordance with an example of the present disclosure;

Figure 3C shows an example vector of three values $\bar{\tilde{z}}_1, \bar{\tilde{z}}_2, \bar{\tilde{z}}_3$ in the signal space to which the soft quantized signal values $\tilde{z}_1, \tilde{z}_2, \tilde{z}_3$ have been hard quantized by a hard quantization module

that are transmittable by a transmitter over the communication channel in accordance with an example of the present disclosure;

Figure 3D shows an example vector of three noise-affected values $\hat{z}_1, \hat{z}_2, \hat{z}_3$ received at a receiver corresponding to a noise-affected version of a vector of hard quantized signal values transmitted over the noisy communication channel by a transmitter, the noise-affected values $\hat{z}_1, \hat{z}_2, \hat{z}_3$ being provided as a channel output vector to a decoder input layer of a decoder neural network for reconstructing a representation of the input data in accordance with an example of the present disclosure;

Figure 4 shows a communication system for conveying data from an information source across a communication channel using joint source and channel coding, illustrating the layers of the encoder and decoder neural network, and the processing of the vector of signal values by the soft and hard quantization modules, in accordance with an example of the present disclosure;

Figure 5 shows an illustration of the forward and backward passes of a training time process for updating the connecting node weights of the encoder and decoder neural networks to minimise an objective function characterising a reconstruction error between input-output pairs of training data in accordance with an example of the present disclosure;

Figure 6 shows an example arrangement of neural network layers for the encoder and decoder neural networks in accordance with an example of the present disclosure;

Figure 7 shows an example training time method for the encoder and decoder neural networks in accordance with an example of the present disclosure;

Figure 8 shows an example run time method for the transmitter and the encoder neural network in accordance with an example of the present disclosure; and

Figure 9 shows an example run time method for the receiver and the decoder neural network in accordance with an example of the present disclosure.

DETAILED DESCRIPTION

[0045] Hereinafter, embodiments of the disclosure are described with reference to the accompanying drawings. However, it should be appreciated that the disclosure is not limited to the embodiments, and all changes and/or equivalents or replacements thereto also belong to the scope of the disclosure. The same or similar reference denotations may be used to refer to the same or similar elements throughout the specification and the drawings.

[0046] As used herein, the terms “have,” “may have,” “include,” or “may include” a feature (e.g., a number, function, operation, or a component such as a part) indicate the existence of the feature and do not exclude the existence of other features.

[0047] As used herein, the terms “A or B,” “at least one of A and/or B,” or “one or more of A and/or B” may include all possible combinations of A and B. For example, “A or B,” “at least one of A and B,” “at least one of A or B” may indicate all of (1) including at least one A, (2) including at least one B, or (3) including at least one A and at least one B.

[0048] As used herein, the terms “first” and “second” may modify various components regardless of importance and do not limit the components. These terms are only used to distinguish one component from another. For example, a first user device and a second user device may indicate different user devices from each other regardless of the order or importance of the devices. For example, a first component may be denoted a second component, and vice versa without departing from the scope of the disclosure.

[0049] It will be understood that when an element (e.g., a first element) is referred to as being (operatively or communicatively) “coupled with/to,” or “connected with/to” another element (e.g., a second element), it can be coupled or connected with/to the other element directly or via a third element. In contrast, it will be understood that when an element (e.g., a first element) is referred to as being “directly coupled with/to” or “directly connected with/to” another element (e.g., a second element), no other element (e.g., a third element) intervenes between the element and the other element.

[0050] As used herein, the terms “configured (or set) to” may be interchangeably used with the terms “suitable for,” “having the capacity to,” “designed to,” “adapted to,” “made to,” or “capable of” depending on circumstances. The term “configured (or set) to” does not essentially mean “specifically designed in hardware to.” Rather, the term “configured to” may mean that a device can perform an operation together with another device or parts.

[0051] For example, the term “processor configured (or set) to perform A, B, and C” may mean a generic-purpose processor (e.g., a CPU or application processor) that may perform the operations by executing one or more software programs stored in a memory device or a dedicated processor (e.g., an embedded processor) for performing the operations.

[0052] The terms as used herein are provided merely to describe some embodiments thereof, but not to limit the scope of other embodiments of the disclosure. It is to be understood that the singular forms “a,” “'an,” and “the” include plural references unless the context clearly dictates otherwise. All terms including technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which the embodiments of the disclosure belong. It will be further understood that terms, such as those defined in commonly used dictionaries, should be interpreted as having a meaning that is consistent with their meaning in the context of the relevant art and will not be interpreted in an idealized or overly formal sense unless expressly so defined herein. In some cases, the terms defined herein may be interpreted to exclude embodiments of the disclosure.

[0053] Figure 1 shows a communication system 100 comprising a transmitter 110 for conveying data across a communication channel 120 to a receiver 130 using joint source and channel coding in accordance with an example of the present disclosure.

[0054] The transmitter 110 and receiver 130 may each be part of respective electronic devices. For example, the electronic device coupled to the transmitter 110 or receiver 130 may be a smartphone, a tablet, a personal computer such as a desktop computer, a laptop computer, a netbook computer, a workstation, a server, a wearable device such as a smart watch, smart glasses, a head-mounted device or smart clothes, an airborne or land drone, a robot or other autonomous device such as industrial or home robots, a smart home appliance such as a television, a smart home device, a media player, a refrigerator, an air conditioner, a cleaner, an oven, a microwave oven, a washer, a drier, an air cleaner, a set-top box, a home automation control panel, a security control panel, a gaming console, a security camera, a microphone, or an Internet of Things device for sensing or monitoring, such as a smart meter, various sensors, an electric or gas meter, a medical device such as a portable medical measuring device, a blood sugar measuring device, a heartbeat measuring device, or a body temperature measuring device, a navigation device, a global positioning system (GPS) receiver, an event data recorder (EDR), a flight data recorder (FDR), avionics, point of sale devices.

[0055] The transmitter 110 includes an information source 111 which may be a source of data originally generated locally to the transmitter (such as an image or video captured by a camera coupled to the transmitter, such as in the electronic device of which the transmitter is a part) or it may be a source of data stored locally to the transmitter that was generated elsewhere, remotely from the transmitter 110. The information source 111 is to be transmitted over the communication channel 120 by the transmitter 110. The information source 111 may store or generate 'raw' or 'uncompressed' data directly or indirectly representative of characteristics of the information source, to allow faithful reproduction of the information source 111 by a given combination of data processing hardware appropriately configured, for example by software or firmware. The information source 111 may be representative of images, documents, audio or video recordings, sensor data, and so on. The information source 111 is any information source suitable for arranging as a sequence of source symbols or fundamental data elements (for example, bits), such as static files or databases or arranged as a sequence over time, as in a stream of data from a sensor or a video camera.

[0056] The transmitter 110 includes at least one processor 112, memory 113 and a carrier modulator 118 coupled to an antenna 119 for transmitting data over communication channel 120. A bus system (not shown) may be provided which supports communication between at the least one processor 112, memory 113, carrier modulator 118 and antenna 119.

[0057] The processor 112 executes instructions that can be loaded into memory 113. The processor 112 can include any suitable number(s) and type(s) of processors or other devices in any suitable arrangement. Example types of processor 112 include microprocessors, microcontrollers, digital signal processors, field programmable gate arrays and application specific integrated circuits.

[0058] The memory 113 may be provided by any structure(s) capable of storing and facilitating retrieval of information (such as data, program code, and/or other suitable information on a temporary or permanent basis). The memory 113 can represent a random access memory or any other suitable volatile or non-volatile storage device(s). The memory 113 may also contain one or more components or devices supporting longer-term storage of data, such as a ready only memory, hard drive, flash memory, or optical disc, which may store software code for loading into the memory 113 at runtime. In use, the processor 112 and memory 113 provide a Runtime Environment (RTE) 114 in which instructions or code loaded into the memory 113 can be executed by the processor to generate instances of software modules in the Runtime Environment 114.

[0059] The memory 113 comprises instructions which, when executed by the one or more processors 112, cause the one or more processors 112 to instantiate an encoder neural network 115, a soft quantization module 116 and a hard quantization module 117. Together, the encoder neural network 115, soft quantization module 116 and hard quantization module 117 may carry out the runtime method described in Figure 8 for encoding input data from information source 111 to hard quantized signal values in a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the transmitter (using the carrier modulator 118 and antenna 119) over the communication channel 120, the signal values representing a transformed version of the input data (which may be a compressed version).

[0060] The carrier modulator 118 operates to in use encode the in-phase and quadrature components of one or more carriers or subcarriers with signal values provided to the carrier modulator 118 by the hard quantization module 117 using an appropriate modulation technique. Where multiple carriers or subcarriers are encoded simultaneously, a suitable multiplexing technique such as orthogonal Frequency-Division Multiplexing (OFDM) may be used. The encoded carriers are then transmitted by the antenna 119 onto the communication channel 120.

[0061] The carrier modulator 118 and antenna 119 may be of conventional construction and may be configured to encode the carriers/subcarriers with signal values mapped only to one or more finite, fixed sets or 'constellations' of symbols of complex IQ representations, such as by quadrature amplitude modulation (QAM) or binary phase-shift keying (BPSK). For example, the carrier modulator 118 and antenna 119 may be compatible with the 5G New Radio standard

such that the transmittable symbols of IQ values are mapped to the 16-QAM, 64-QAM or 256-QAM constellations. The carrier modulator 118 and antenna 119 may be hard-wired to work only with these symbols, and they may not be able to transmit signal values or symbols that are not within these standard constellation sets.

[0062] The communication channel 120 may be used to convey information from one or more such transmitters 110 to one or more such receivers 130. The communication channel 120 may be a physical connection, e.g., a wire, or a wireless connection such as a radio channel as in the example shown in Figure 1. There is an upper limit to the performance of a communication system 100 which depends on the system specified. In addition, there is also a specific upper limit for all communication systems which no system can exceed. This fundamental upper limit is an upper bound to the maximum achievable rate of reliable communication over a noisy channel and is known as Shannon's capacity.

[0063] The communication channel 120, including the noise associated with such a channel, is modelled and defined by its characteristics and statistical properties. Channel characteristics can be identified by comparing the input and output of the channel (as will be shown below in relation to Figures 3C and 3D), the output of which is likely to be a randomly distorted version of the input. The distortion indicates channel statistics such as additive noise, or other imperfections in the communication medium such as fading or synchronization errors between the transmitter 110 and the receiver 130. Channel characteristics include the distribution model of the channel noise, slow fading and fast fading. Common channel models include binary symmetric channel and additive white Gaussian noise (AWGN) channel.

[0064] The receiver 130 includes at least one processor 132, memory 133 and a carrier demodulator 138 coupled to an antenna 139 for receiving data over communication channel 120. A bus system (not shown) may be provided which supports communication between at the least one processor 132, memory 133, carrier demodulator 138 and antenna 139.

[0065] Similarly to the processor 112 and memory 113 of the transmitter 110, in the receiver 130, the processor 132 executes instructions that can be loaded into memory 133, and in use provide a Runtime Environment (RTE) 134 in which instructions or code loaded into the memory 133 can be executed by the processor to generate instances of software modules in the Runtime Environment 134. The memory 133 comprises instructions which, when executed by the one or more processors 132, cause the one or more processors 132 to instantiate a decoder neural network 135.

[0066] The antenna 139 of the receiver 130 receives a noise-affected version of the source symbols transmitted by the antenna 119 of the transmitter 110, the noise having been added by the communication channel 120. The carrier demodulator 138 demodulates these noisy

symbols, for example, by coherent demodulation, and passes them to the decoder neural network 135.

[0067] These noisy demodulated symbols are then mapped by the decoder neural network 135 to a reconstructed representation of the originally input source data to reconstruct the information source 111. The receiver 130 thus includes an information sink 131 to which the reconstructed representation of the input data decoded by the decoder neural network 135 is provided.

[0068] In accordance with the present disclosure, the encoder neural network 115 and the decoder neural network 135 have been trained together to seek to minimise an objective function characterising a reconstruction error between input-output pairs of training data from the information source 111 passed to the encoder neural network 115 and the representation of the input data reconstructed at the decoder neural network 135 using an appropriate optimization algorithm operating on the objective function.

[0069] In this way, the input data from the information source 111 (such as the image or video) encoded and transmitted by the transmitter 110 can be received and decoded at the receiver 130 to allow a reconstructed representation of the original input image or video to be generated at information sink 131.

[0070] Reference will now be made to Figures 2, 3, 4, 8 and 9 which set out in more detail how the transmitter 110 and receiver 130, and the trained encoder neural network 115 and decoder neural network 135, operate to transmit data from information source 111 across communication channel 120 by joint source and channel coding.

[0071] Neural networks are machine learning models that employ multiple layers of nonlinear units (known as artificial “neurons”) to generate an output from an input. Neural networks may be composed of several layers, each layer formed from nodes. Neural networks can have one or more hidden layers in addition to the input layer and the output layer. The output of each layer is used as the input to the next layer (the next hidden layer or the output layer) in the network. Each layer generates an output from its input using a set of parameters, which are optimized during the training stage. For example, each layer comprises a set of nodes, the nodes having learnable biases and their inputs having learnable weights. Learning algorithms can automatically tune the weights and biases of nodes of a neural network to optimise the output in order to minimise an objective function using an optimisation algorithm such as gradient descent or stochastic gradient descent.

[0072] As shown in Figure 4, in the transmitter 110, the encoder neural network 115 has an encoder input layer 115i having input nodes for receiving a vector x of input data from an information source 111 to be compressed. The encoder neural network 115 also has an

encoder output layer 115o having output nodes for outputting a vector z of values in a signal space for modulating a carrier signal for transmission of a compressed version of the input data across a communication channel. The encoder neural network 115 also has one or more hidden layers 115h having connecting nodes connecting the encoder input layer 115i to the encoder output layer 115o, the connecting nodes having weights to map the input data received at the encoder input layer 115i to provide at the encoder output layer 115h a compressed representation of the input data. The vector z of signal values output by the encoder output layer 115o are quantized first by soft quantization module 116 to produce a soft quantized signal vector $\tilde{z}$ and then hard quantization module 117 to produce a hard quantized signal vector $\bar{z}$ of signal values assigned to symbols in a predetermined finite set S of the alphabet symbols in a constellation set transmittable by the transmitter hardware of the transmitter 110 over the communication channel 120. The hard quantized signal vector $\bar{z}$ is passed to the carrier modulator 118 to modulate a carrier signal for transmission by antenna 119 over the communication channel 120.

[0073] In the receiver 130, antenna 139 receives a noise-affected version of the source symbols transmitted by the antenna 119, and the carrier demodulator 138 demodulates these noisy symbols and passes them to the decoder neural network 135 as a channel output vector $\hat{z}$ of noise-affected version of the vector $\bar{z}$ of hard quantized signal values transmitted over the noisy communication channel 120. The decoder neural network 135 comprises a decoder input layer 135i having nodes corresponding to channel output vector $\hat{z}$, and a decoder output layer 135o having output nodes for outputting as a vector $\hat{x}$ a reconstructed representation of the vector x of input data provided from the information source 111 to the encoder input layer 115i of the encoder neural network 115. The decoder neural network 135 also has one or more hidden layers 135h having connecting nodes connecting the decoder input layer 135i to the decoder output layer 135o, the connecting nodes having weights to map the channel output vector $\hat{z}$ of noise-affected hard quantized signal values received at the decoder input layer 135i to provide at the decoder output layer the reconstructed representation of the input data as vector $\hat{x}$.

[0074] The noisy demodulated symbols $\hat{z}$ are thus mapped by the decoder neural network 135 to a reconstructed representation $\hat{x}$ of the originally input source data x to reconstruct the information source 111. The receiver 130 thus includes an information sink 131 to which the reconstructed representation of the input data $\hat{x}$ decoded by the decoder neural network 135 is provided.

[0075] In the sense that the encoder neural network 115 and decoder neural network 135 are configured to provide at the decoder output layer 135o a reconstructed representation $\hat{x}$ of the originally input source data x provided to the encoder input layer 115i of the encoder neural

network 115, they can be said to together provide an autoencoder, with the communication channel in between considered as an untrainable layer. An autoencoder is a neural network that is trained to copy its input (in this case the information source 111) to its output at the information sink 131. Autoencoders can provide dimensionality reduction at an intermediate layer (in this case the encoder output layer 115o/decoder input layer 135i) arranged to have fewer nodes than the encoder input layer 115i. In this respect, the encoder output layer 115o of the encoder neural network 115 may have a lower dimension than the encoder input layer 115i, so that a low-dimensional representation of the input data is learned, which may represent a compression of the input data. Generally, however, the output dimension of the encoder neural network is a parameter specified when the architecture of the encoder neural network is specified by the system designed. The number of nodes of the encoder output layer 115o is generally chosen to correspond to the channel bandwidth available on the communication channel 120 to transmit the encoded data from the information source 111. If a lot of bandwidth is available for use in the communication channel 120, then the number of nodes of the encoder output layer 115o may actually be chosen to be a value larger than the number of nodes in the encoder input layer 115i. This does not mean that the output of the encoder neural network 115 will convey more information than the input. The encoder output is limited to a specified constellation, whereas the input can take any value. In this way, the encoder learns an optimum transformation to map the input information source to the channel, that allows a faithful reconstruction of the information source at the decoder.

[0076] Where the output of the encoder output layer 115o is configured to be a complex representation of IQ signal values for carrier modulation at carrier modulator 118 for transmission over the communication channel 120, the encoder 115 is trained to perform joint source and channel coding in a single shot from the input source data x to the transmittable signal values z (later quantized to $\vec{z}$). In this way, the encoder neural network of the autoencoder can replace the traditional source encoder and channel encoder. Similarly, the decoder neural network 135 performs joint source and channel decoding from the noisy demodulated symbols $\hat{z}$ to a reconstructed representation $\hat{\mathrm{x}}$ of the input data x in a single shot. In this way, the decoder neural network of the autoencoder can replace the traditional source decoder and channel decoder. Thus, the encoder and decoder neural networks can be used as a form of joint source channel coding. This provides simplicity over having two separate systems for encoding and two separate systems for decoding. However, the limitations of the ML-optimisation approaches being incompatible with existing hardware that can transmit only fixed constellation sets are overcome as the encoder neural network 115 and decoder neural network 135 used for joint source channel coding in accordance with examples of the present disclosure can be trained for use with existing transmitter technology transmitting different fixed constellations, as well as being trained for different channels and different input sources. Thus,

using neural networks for conveying information across a channel enables a high degree of freedom and a high degree of compatibility.

[0077] In an autoencoder, the goal of the training of the node weights is to make sure that the input data x provided to the encoder input layer 115i can be reconstructed as $\hat{\mathrm{x}}$ by the decoder neural network 135 based on the encoder output z that is received and decoded by the decoder neural network 135. In the present case, to ensure compatibility with common communication hardware, the encoder output z of values that may in principle be unconstrained to take any value in the signal space is hard quantized to a vector $\bar{z}$ of hard quantized signal values corresponding to an alphabet S of values a constellation set transmittable by the transmitter 110. That is, the encoder output values of the vector z are not limited to taking values in a particular constellation set. Thus, what is passed to the decoder neural network 135 is not a noise-affected version of the unconstrained values z output by the encoder neural network 115, but a noise-affected version of the vector $\bar{z}$ of hard quantized signal values produced by the soft quantization module 116 and hard quantization module 117 and transmitted by carrier modulator 118 and antenna 119 of transmitter 110.

[0078] Thus, in the training process (explained below in more detail with reference to Figures 5 and 7), the encoder neural network 115 and decoder neural network 135, update their weights to learn to minimise an objective function characterising a reconstruction error between input-output pairs of training data from the information source 111 passed to the encoder neural network 115 and the representation of the input data reconstructed at the decoder output layer 135o of the decoder neural network 135 using an appropriate optimization algorithm operating on the objective function, wherein the unconstrained output of the encoder neural network 115 is subsequently quantized to a fixed constellation set and the decoder neural network receives a noise-affected version of this quantized output. In this way, the encoder neural network 115 and decoder neural network 135 provide an autoencoder in which the compressed signal is adapted to be assigned to a fixed signal alphabet or constellation set transmittable using common transmission hardware of transmitter 110.

[0079] The use of the soft quantization module 116 to provide a differentiable representation of the quantized signal for transmission is key to allowing the encoder neural network 115 and decoder neural network 135 to learn to minimise reconstruction errors in the communication system in which a fixed signal alphabet or constellation set of discrete signal values as encoded symbols for transmission is used. The differentiable output of the soft quantization module 116 is usable to identify the error in the input-output pairs during training and to update the weights using, for example, gradient descent wherein the gradient of the objective function characterising the error may be calculated in a backward pass using the chain rule and by automatic differentiation of the function stack used by the processor (which may or may not be

the processors 112 or 132, as the training may occur using different computing hardware) to calculate the decoder output $\hat{x}$ from the encoder input x in the forward pass.

[0080] Thus, by including the channel in the training process, by the decoder neural network 135 receiving a noise-affected version $\hat{z}$ of the quantized output $\bar{z}$ in the forward pass, the encoder neural network 115 and decoder neural network 135 are trained to undo the effect of the channel on the signal, such as the effect of the channel noise of the signal. The encoder neural network 115 and decoder neural network 135 can also be trained to learn a representation of the source data 111 and recover the input with the highest fidelity possible. Further, as the neural networks are trained to optimise the mapping, the efficiency of the transmission of the information source over the communication channel can be greater than in separate source and channel coding, where the separate removal and subsequent addition of redundancy may be inherently inefficient for transferring the data.

[0081] In a detailed example of a forward pass at runtime, as shown in Figure 2, the information source 111 in the example is an image of a boat, which may be an uncompressed or raw bit map. Referring to Figure 8, which shows an example run time method 800 for the transmitter 110, although this is not a required step of the method of the present disclosure, in step 801, a vector of input data from the information source 111 in the form of bit symbols x_1, x_2, x_3... x_n where $x \in \mathbb{R}^n$ is received at input nodes in an encoder input layer 115i for compression by the encoder neural network 115.

[0082] In step 803, the vector of input data x_1, x_2, x_3, ... , x_n received at the encoder input layer 115i is mapped to provide at the encoder output layer 115o a compressed representation of the input data. The mapping is performed by one or more hidden layers 115h of the encoder neural network 115 having connecting nodes connecting the encoder input layer 115i to the encoder output layer 115o. The connecting nodes have weights to map the input data x_1, x_2, x_3, ... , x_n to a compressed representation of the input data in a latent vector z of values z_1, z_2, z_3, ... , z_k in the signal space, where the signal space is a complex representation of the in phase and quadrature components for modulating one or more carriers in the communication channel 120. Thus, the connecting nodes of the hidden layers 115h map the input vector x_1, x_2, x_3, ... , x_n where $x \in \mathbb{R}^n$ to an output vector z_1, z_2, z_3, ... , z_k where $z \in \mathbb{C}^k$ of k channel inputs using a nonlinear encoder function $f_\theta(x)$, where k < n to provide the compressed representation. Figure 3A shows an example vector of three values, z_1, z_2, z_3 each having in-phase and quadrature components in an IQ signal space for modulating a carrier signal for transmission across a communication channel to which a vector of input data from an information source has been mapped by an encoder neural network in accordance with an example of the present disclosure. As illustrated by the gradient map of potential values for z_1, z_2, z_3, ... , z_k in the IQ signal space in Figure 3A, the output z of the encoder neural network 115 is unconstrained and

can take a range of values that are subsequently assigned through the soft quantization module 116 and hard quantization module 117 to specific IQ values in a constellation set. It should be noted that the gradient map showing in Figure 3A an illustration of the distribution of likely signal values for the outputs for $z_1, z_2, z_3, \ldots, z_k$ in the IQ signal space is exaggerated for ease of understanding, and through training in conjunction with the quantization modules the encoder neural network 115 will tend to learn to output values at or very close to the constellation set for transmission. Nevertheless, as can be seen in Figure 3A, the values of the encoder output layer 115o vector elements z_1, z_2, z_3 are near to but away from the exact signal values for symbols in the constellation set S of the 16QAM constellation transmittable by the transmitter 110, to which the signal values are to be quantized.

[0083] In step 805, the compressed representation of the input data $z_1, z_2, z_3, \ldots, z_k$ taking unconstrained output signal space values is quantized to the exact signal values for symbols in the constellation set S of the constellation transmittable by the transmitter 110. This happens in two stages.

[0084] Firstly, the soft quantization module 116 operates a differentiable function $q(z_i)$ over the signal space that transforms the unconstrained output signal space values z_i to a smoothed approximation $\tilde{z}_i$ of a signal value in a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the transmitter over the communication channel. That is, given a finite set of constellation points $S = \{s_1, s_2, s_3 \ldots s_m\}$ for each latent element z_i we calculate its approximate quantized version as:

$$\tilde{z}_i = q(z_i) = \sum_{j=1}^{m} \frac{e^{-\sigma d_{ij}}}{\sum_{k=1}^{m} e^{-\sigma d_{ik}}} s_j =$$

[0085] where σ is a parameter controlling the "hardness" of the assignment, and $d_{ij} = \left\|z_i - s_j\right\|^2$ is the squared Euclidean distance between complex latent element z_i and constellation point s_j. In general, soft quantization produces latent variables that take values outside the constellation S.

[0086] As the vector $\tilde{z}_i$ is generated by a smoothed function $q(z_i)$ to provide an approximation of the signal values in a predetermined finite set S of symbols of a predefined constellation, the function $q(z_i)$ is differentiable and usable to identify the error in the input-output pairs during training and to update the weights using, for example, gradient descent wherein the gradient of the objective function characterising the error may be calculated in a backward pass using the chain rule and by automatic differentiation of the function stack used by the processor.

[0087] As can be seen from Figure 3B, the vector elements z_1, z_2, z_3 input to the soft quantization module 116 have been transformed by the smoothed function $q(z_i)$ to provide vector elements $\tilde{z}_1, \tilde{z}_2, \tilde{z}_3$ of vector $\tilde{z}$ that are nearer to but still away from the exact signal values for symbols in the constellation set $\mathcal{S}$ of the 16QAM constellation transmittable by the transmitter 110, to which the signal values are to be quantized. Even for large values of parameter σ, there is still a gap between the soft quantized values $\tilde{z}$ and the exact signal values of the constellation for transmission. This is shown by the gradient map of the distribution of values taken by $\tilde{z}$ in Figure 3B showing a spread of signal values around the 16QAM constellation points. In general, the signal values $\tilde{z}$, being away from the standard 16QAM values, would not be compatible for transmission by standard communication hardware.

[0088] To force the transmitter to provide transmittable signal values, in a second stage of quantization, the approximated signal values $\tilde{z}$ output by the soft quantization module 116 are passed to a hard quantization module 117 that in use transforms them exactly to the nearest signal value assigned to the symbol in the predetermined finite set S of the alphabet by:

$$\bar{z} = [\tilde{z}_i] = s_j$$

where $j = \arg\min d_{ik}$ and $k \in \{1, \ldots m\}$.

[0089] As can be seen in Figure 3C, the vector elements $\tilde{z}_1, \tilde{z}_2, \tilde{z}_3$ of vector $\tilde{z}$ have been assigned by the hard quantization module 117 exactly to the signal values $\bar{z}_1, \bar{z}_2, \bar{z}_3$ assigned to the nearest symbols of the set S of symbols making up the alphabet of the standard 16QAM constellation, as shown by the 16 points laid out in the IQ signal plot. In the Figure 3C, it can be seen that the elements of the vector $\bar{z}$ output by the hard quantization module 117 take no values other than those of the standard 16QAM constellation. The hard quantized signal values $\bar{z}_1, \bar{z}_2, \bar{z}_3 \ldots \bar{z}_k$ represent the compressed version of the input data $x_1, x_2, x_3 \ldots x_n$.

[0090] In step 807, the hard quantized signal values $\bar{z}_1, \bar{z}_2, \bar{z}_3 \ldots \bar{z}_k$ are passed as a vector from the hard quantization module 117 to the carrier modulator 118 to modulate one or more carrier signals, for example, by OFDM, for transmission by antenna 119 over the communication channel 120. There the runtime process 800 at the transmitter 110 for transmitting of the input data $x_1, x_2, x_3 \ldots x_n$ ends, and is repeated for all the data for the information source 110 until the information source (in the example the image of the boat) has been transmitted.

[0091] Referring now to Figure 9, which shows an example run time method 900 for the receiver and the decoder neural network in accordance with an example of the present disclosure, at the receiver 130, in step 901 the one or more carrier signals are received from communication channel 120 by the antenna 139, and the one or more carrier signals are

demodulated by the carrier demodulator 138, for example by coherent demodulation, to recover a noise-affected version $\hat{z}$ of the hard quantized signal values $\bar{z}$ of the source symbols transmitted by the antenna 119 of the transmitter 110. As can be seen in Figure 3D, which shows a distribution of demodulated noise-affected signal values $\hat{z}$ received at the receiver 130, which shows that the noise in the channel leads to a distribution of different detected signal values $\hat{z}$ different to the hard quantized signal values $\bar{z}$. The spread of detected signal values $\hat{z}$ is a randomly distorted version of the input based on a noise distribution based on the channel characteristics which may include additive noise, or other imperfections in the communication medium such as fading, multipath propagation, interference or other attenuation. The noise in the channel can be modelled as, for example, an additive white Gaussian noise (AWGN) channel model, which, as will be explained below, may be useful in training the encoder and decoder neural networks, rather than having to use empirical data for transmitted input-output pairs.

[0092] In the example shown in Figure 3D, the vector $\hat{z}$ of demodulated carrier signals received at a receiver 130 takes noise-affected values $\hat{z}_1, \hat{z}_2, \hat{z}_3$ different to (i.e. away from) the corresponding vector of hard quantized signal values $\bar{z}_1, \bar{z}_2, \bar{z}_3$ (shown in Figure 3C) transmitted over the noisy communication channel by a transmitter 110, and also away from the set S of symbol values making up the constellation shown in Figure 3C.

[0093] The noise-affected values $\hat{z}_1, \hat{z}_2, \hat{z}_3$ are then provided as a channel output vector $\hat{z}$ to a decoder input layer 135i of a decoder neural network 135 for reconstructing a representation of the vector of input data x in accordance with an example of the present disclosure.

[0094] In step 903, the hidden layers 135h of the decoder neural network 135 map the vector $\hat{z}$ of noise-affected values to a reconstructed representation $\hat{\mathrm{x}}$ of the originally input source data x to reconstruct the information source 111 in the information sink 131. In this respect, the hidden layers 135h apply a non-linear decoder function $g_{\theta\varphi}(\hat{z})$ to produce a reconstruction $\hat{\mathrm{x}}$ of the input, where $\hat{\mathrm{x}} \in \mathbb{R}^{\mathrm{n}}$.

[0095] Thus as can be seen in Figure 2, vectors of input data x from the image of the boat from the image source 111 are mapped by the encoder neural network 115, soft quantization module 116 and hard quantization module 117 of the transmitter 110 to a vector $\bar{z}$ of hard quantized signal values which is provided as an input 118i to the carrier modulator 118 and transmitted over the communication channel 120 by antenna 119. At the receiver 130, this is received and demodulated as an output 138o of the carrier demodulator 138 as a vector $\hat{z}$ of noise-affected signal values which is then mapped by the decoder neural network 135 to a reconstruction $\hat{\mathrm{x}}$ of the input data allowing.

[0096] The connecting node weights of the hidden layers 115h and 135h of the encoder neural network 115 and decoder neural network 135 have been trained together to seek to minimise an objective function characterising a reconstruction error between input-output pairs of training data from the information source 111 passed to the encoder neural network 115 and the representation of the input data reconstructed at the decoder output layer 135o of the decoder neural network 135 using an appropriate optimization algorithm operating on the objective function.

[0097] In this way the, encoder neural network 115 and decoder neural network 135 learn to optimise the compression of the data of the image of the boat and its transmission as carrier symbols of a fixed constellation set over a noisy communication channel such that the image of the boat is reconstructed at the information sink 131 of the receiver with few errors.

[0098] The training time process for optimising the weights of the encoder neural network 115 and decoder neural network 135 to minimise reconstruction errors will now be described with further reference to Figures 5 and 7.

[0099] Once all the transformations blocks of the communication system as shown in Figure 5 have been designed and initialised with suitable initial encoder and decoder weights and parameters θ, φ, the weights of the encoder neural network 115 and decoder neural network 135 are jointly optimized end-to-end in an unsupervised manner by passing training data sample vectors x each having elements $x_1, x_2, x_3, \ldots x_n$ as inputs through the communication system 100 (or a simulation thereof using a channel model to add noise) and receiving its reconstruction vector $\hat{\mathrm{x}}$ having elements $\hat{x}_1, \hat{x}_2, \hat{x}_3, \ldots \hat{x}_n$ in a forward pass of training data through the transformation blocks (as indicated by the solid line). That is, in step 701 the vectors $(\mathrm{x}, \hat{\mathrm{x}})$ are received and form input-output pairs of a set of training data from the information source 111 passed to the encoder neural network 115

[00100] In examples, in the training phase, the input-output pairs of vectors $(\mathrm{x}, \hat{\mathrm{x}})$ of training data may be calculated empirically, by the transmitter 110, in the forward pass, encoding and transmitting the encoded compressed version $\bar{z}$ of the input vector x of training data across the communication channel 120 where the signal values are subsequently received and decoded by receiver 130. In this way, the encoder and decoder neural networks can be optimised to take into account the noise in the channel through training based on a empirical data capturing the effects of channel noise on the transmission.

[00101] In other examples, as shown in the Figure 5 example, in the training phase, the input-output pairs of vectors $(\mathrm{x}, \hat{\mathrm{x}})$ of training data may be generated using a model of the communication channel 120 to estimate channel noise and add it to the transmitted hard quantized signal values $\bar{z}$ to generate a noise-affected version $\hat{z}$ of the vector of hard quantized

signal values for subsequent decoding and reconstructing of the output training data vector $\hat{\mathrm{x}}$ by the decoder neural network 135. In these examples, a channel model can be adopted that simulates the practical channel experienced in the operational regime. For simplicity, a complex additive white Gaussian noise (AWGN) channel model can be adopted, which produces the channel output $\hat{z} = \bar{z} + \eta$, where $\eta \in \mathbb{C}^k$ is a vector containing elements drawn from zero-mean Gaussian distribution of variance σ^2. However, in general, the channel model can be any model that simulates an arbitrary transformation of the hard quantized channel input vector $\bar{z}$ transmitted by the transmitter 110.

[00102] The training process may perform batchwise optimisation across groups of input-output pairs, such as using gradient descent to find a gradient error in the forward pass and determine an update to the weights. In other examples, stochastic gradient descent may be used in which the error is determined and weights updated for each input-output pair of vectors $(\mathrm{x}, \hat{\mathrm{x}})$ of training data, before the next of pair of vectors $(\mathrm{x}, \hat{\mathrm{x}})$ of the training data are determined using the updated weights.

[00103] When a batch or a single input-output pair of vectors $(\mathrm{x}, \hat{\mathrm{x}})$ of training data have been received to optimise and update the weights in the training process, in step 703, an objective function is determined characterising a reconstruction error between the input-output pairs of vectors $(\mathrm{x}, \hat{\mathrm{x}})$ of training data. In the example, as shown in Figure 5, the reconstruction error for the objective function is characterised using the Mean Squared Error loss between x and $\hat{x}$, calculated by:

$$\mathrm{MSE}(\mathrm{x}, \hat{\mathrm{x}}) = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \hat{\mathrm{x}}_\mathrm{i})^2$$

[00104] Other objective functions characterising the reconstruction error may be used.

[00105] Once the objective function has been calculated for the input-output pair of vectors $(\mathrm{x}, \hat{\mathrm{x}})$ of training data, the method further comprises, in steps 705 and 707 which may be performed together, using an appropriate optimisation algorithm operating on the objective function, updating the connecting node weights of the hidden layers of the encoder neural network and decoder neural network to seek to minimise the objective function.

[00106] In step 705, the gradient descent optimisation algorithm is used to seek to minimize the objective function by using a differential of the objective function to determine the gradient and the direction towards a minimum value for the objective function. Even though the output $\bar{z}$ of the hard quantization module 117 is non-differentiable, a differentiable path through the transformation blocks to determine a gradient in the MSE objective function is provided by the

soft quantization module 116 which provides a smoothed approximation $\tilde{z}$ of the signal values in the constellation set.

[00107] Thus, as can be seen in Figure 5, in a backward pass through the communication system (as indicated by the dotted line), the gradient descent algorithm operates on the objective function based on a differential of at least the decoder neural network 135, the soft quantization module 116 and the encoder neural network 115. For example, using backpropagation, the gradient of the objective function can be efficiently calculated with respect to the weights in the decoder neural network 135 and encoder neural network 115, for example by unstacking the elementary functions used to compute the forward pass, and by repeatedly applying the chain rule to autodifferentiate them and determine the gradient with respect to the weights in the encoder neural network 115 and decoder neural network 135 by backpropagation.

[00108] Once the gradient of the object function is determined with respect to the weights in the encoder neural network 115 and decoder neural network 135, in step 707, the connecting node weights of the hidden layers of the encoder neural network and decoder neural network are updated to seek to minimise the objective function. In examples, this achieved in the gradient descent optimisation method by the using the determined gradient to estimate an update to the weights of the encoder neural network 115 and decoder neural network 135 that is expected to step the objective function towards a minimum, where the local gradient is zero.

[00109] Once the estimate of the update to the weights is determined by optimisation method, the weights of the encoder neural network 115 and decoder neural network 135 are updated and in step 709, it is checked whether there are more samples of training data in the training set, and, if there are, the process returns to step 701 and the next batch or training sample is received and the optimisation method is carried out again to further optimise the weights of the neural networks. If training over the training set is complete, the process ends and a trained encoder neural network 115 and decoder neural network 135 are provided for use in an operational communication system 100 for transmitting input data over a communication channel 120.

[00110] The separation of quantization into soft and hard steps deals with the issue of non-differentiability of the quantization operation, and enables the end-to-end learning of all the transformation blocks of the communication system 100 through gradient descent. The soft quantization is a differentiable approximation of the hard quantization function, where $\tilde{z}_i$ is a combination of symbols s_j. As only one symbol in S should be chosen (instead of a combination), the hard quantization selects the main component of $\tilde{z}_i$, so that $\overline{z}_i$ belongs to S. The hard quantization is not a differentiable operation; however, as it is applied after the soft

quantization approximation, we have $\overline{z}_i \sim \tilde{z}_i$, enabling the system to be differentiable while guaranteeing that the output vector $\overline{z}$ takes values from set S.

[00111] For the optimization process, as the encoder and decoder blocks are built as artificial neural networks with learnable parameters so that the transformation from/to data to latent representation (code) can be learned directly from data. If the constellation symbols S are predefined, as it is the case when using standard communication hardware and protocols, the soft and hard quantization blocks act as constraints for the optimization and the objective function. In this way, the objective function, being differentiable despite the generation of discrete symbol values in a constellation, allows the gradient in the objective function to be evaluated, allowing the system to learn weights of the encoder neural network 115 and decoder neural network 135 to minimise the reconstruction error.

[00112] If a channel model is used in the forward pass of the training process, rather than empirically generating training data, the channel model can be included directly in the backward pass in the optimisation algorithm. If the channel model used is differentiable, it can be used directly in the backpropagation stage. If it is not differentiable, a generative adversarial network (GAN) may be used to learn a differentiable representation of the channel model. In this way, the encoder and decoder neural networks can be optimised to take into account the noise in the channel through training based on a theoretical noise model of the communication channel.

[00113] It should be noted that, in other examples, the objective function may characterise and optimise against further constraints and characteristics of the system, such as to obtain an average power in the symbols transmitted across the communication system 100, so as to ensure the learned coding system satisfies an average power constraint.

[00114] That is, another constraint of existing communication systems (aside from them permitting only a fixed constellation of transmission symbols) is that the transmitted message should satisfy an average power constraint P. To satisfy this constraint, the constellation S is designed so that, assuming all symbols have the same probability of occurrence, the power is equal or smaller than P. Then, to ensure the communication system 100 satisfies this power constraint, the encoder neural network 115 and decoder neural network 135 are trained to learn a joint source and channel coding scheme that results in the symbols being used with equal probability.

[00115] In examples, this is achieved by adapting the objective function to add to the component characterising the reconstruction error (in the examples, the Mean Square Error) an additional component that characterises a relative entropy between the probability distribution of the symbols $\overline{z}$ of the hard quantized signal values output from the encoder neural network 115 and the hard quantization module 117, and a uniform distribution across the symbols s_j of

the constellation set S. Thus, as a result of the optimization process for the objective function, encoder neural network 115 and decoder neural network 135 of the communication system 100 learn a joint source and channel coding such that the probability distribution of the transmitted hard quantized symbols $\bar{z}$ (based on the calculation of the soft quantized symbols $\tilde{z}$) is such that an average power constraint P on the communication channel is satisfied. This also results in the available fixed constellation points being used in the most optimal manner.

[00116] In examples of the present disclosure, the Kullback-Liebler (KL) divergence between the symbol distribution and the uniform distribution is used in the objective function to characterise the relative entropy for minimisation during training. That is, the uniform distribution is one on which every symbol s_j is transmitted with equal probability across the communication channel 120. Specifically, the KL Divergence term in the objective function assesses the difference between the distribution of the soft-quantized latent variables $\tilde{z}$ and uniform distribution U over the constellation points s_j. To model the distribution of the soft-quantized latent variables $\tilde{z}$, softmax probabilities are utilised, which are calculated by the soft quantization module. Specifically, the average probabilities for all latent elements $\tilde{z}$ are calculated, and their KL Divergence from discrete uniform distribution are then calculated.

[00117] In this example then, the objective function to be minimised in the training process is expressed as:

$$\mathrm{L} = \mathrm{MSE}(\mathrm{x}, \hat{\mathrm{x}}) + \lambda \mathrm{D}_{\mathrm{KL}}(\mathrm{p}_{\tilde{z}} \,||\, \mathrm{U})$$

where λ is a weight parameter.

[00118] In the examples above, the training optimises the transmission on a predefined constellation (e.g., 16QAM). In other examples, the training process can be adapted to learn which one of a set of available predefined constellations optimizes the performance.

[00119] In these examples, the symbol alphabet set size S is a parameter of the soft and hard quantization modules. During training the gradient descent algorithm is further operated over a range of symbol alphabet set sizes S of different cardinality (i.e., different constellations are learned) to learn the optimal modulation scheme for the information source. In this way, the optimal transmittable constellation can be learned for transmitting the information source across the communication channel with minimum distortion.

[00120] Thus, after being defined through training, the encoder neural network 115 and decoder neural network 135 of the communication system 100 can been optimized to: (a) encode the input data through joint source and channel coding into a fixed constellation of symbols, which may have been selected from different available constellation sets to be the

most optimal; (b) create a coded and compressed representation of the input data as a channel input signal in the constellation sets, the coding being resilient to channel noise; (c) satisfy an average power constraint of transmitted signals; (d) decode noisy received symbols back into an uncompressed reconstruction of the input data.

[00121] The overall communication system 100 in accordance with the examples is greatly simplified compared to prior art separate source coding and channel coding in terms of the number of steps taken. Moreover, the communication system 100 provides robust communication that is resilient to the effect of noise on the transmitted signal, and can operate reliably in scenarios with varying channel conditions.

[00122] An example suitable architecture of the layer structure for the encoder neural network 115 and decoder neural network 135 for use in wireless image transmission is shown in Figure 6. As can be seen, the architecture includes a convolutional encoder 115 and decoder 135, made of sub-blocks. For the encoder 115, each sub-block contains a convolutional layer, followed by a generalized divisive normalization (GDN) layer and parametric rectified linear unit (PReLU) activation. For the decoder 135, each block is built of a single transposed convolution, inverse GDN and PReLU activation. In the encoder 115, 4 sub-blocks are used followed by a convolutional layer and GDN. The decoder 135 mirrors the architecture of the encoder 115 and 4 sub-blocks are followed by transposed convolution and IGDN layer. To ensure the input x and output $\hat{x}$ are within the same range, a sigmoid activation is applied after the last IGDN layer in the decoder, which effectively maps unconstrained values of the output to [0, 1] range.

[00123] It should be noted that, in accordance with the present disclosure, the communication channel 120 should be understood as any transformation from the channel input space to the channel output space that includes a random transformation due to the channel. This may include additive noise, interference, or other stochastic properties of the channel that will randomly transform the transmitted signal, e.g., fading and multi-path effects in wireless channels. Thus, the reference to the noise-affected version $\hat{z}$ of the of the vector of hard quantized signal values $\bar{z}$ received at the decoder should be understood to indicate that the input $\hat{z}$ to the decoder is a vector of values correlated with the transmitted vector $\bar{z}$ of signal values (which is itself correlated with the input data x from the information source), transformed by the communication channel 120, whether that transformation is 'noise' or another random channel transformation.

[00124] In this respect, in accordance with the present disclosure, the communication channel 120 should be understood as encompassing any channel that has, in an input signal space, a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal

values transmittable by the transmitter over the communication channel (which applies a random transformation to the channel output space).

[00125] Thus, although the examples described above disclose the transmitted vector $\bar{z}$ of signal values being raw signal values in IQ space corresponding to the constellation transmittable by the transmitter over the communication channel 120, in other embodiments the communication channel 120 may be one that also includes an existing channel encoder and decoder scheme, in which the signal space of the channel input may be the predetermined finite set symbols of the channel code (which could be bit values) for modulating a signal carrier for providing input signals in the alphabet in the input signal space for the communication channel. Thus, besides random noise applied by the communication channel, the transformation applied by the communication channel may, in embodiments, also include an existing channel code. Thus, in these embodiments, the predefined alphabet of symbols for quantization may be a message alphabet for an existing channel code by which the input signals to the given communications channel are modulated. In this case, the hard quantized channel inputs $\bar{z}$ will be mapped into the message alphabet of the corresponding channel code (rather than, for example, the raw IQ values of a transmittable constellation, as in the embodiments described above). The noise-affected channel output $\hat{z}$ input to the decoder neural network 135 may correspond to the hard-decoded message of the channel decoder. In this respect, in these embodiments, the encoder neural network 115 and decoder neural network 135 may learn an optimum mapping of the input information source 111 to inputs of an existing channel code of the communications channel 120 that reduces reconstruction errors at the output 131 of the decoder neural network135. Although acting as an outer code in these embodiments, this learned coding of the encoder neural network 115 and decoder neural network 135 is still optimised based on the characteristics of the communication channel 120 to reduce reconstruction errors, even though in these alternative embodiments the communication channel 120 includes an existing channel code. This is unlike existing modular source codes which are defined independently of the random transformation applied by any channel.

[00126] Further, in accordance with the present disclosure, it should be understood that the dimension of the vector $\bar{z}$ of hard-quantized signal values that are transmitted over the communication channel need not be equal to the dimension of the channel output vector $\hat{z}$ corresponding to the noise-affected version of the vector of hard quantized signal values transmitted over the noisy communication channel. That is, in embodiments the version $\hat{z}$ of the vector of hard quantized signal values transmitted over the noisy communication channel that is input into the decoder input layer may have a dimensionality different to the transmitted vector $\bar{z}$ of hard quantized signal values. This may be due, for example, to characteristics of the receiver, the detection and demodulation process revealing a different signal vector. For

example, this may be due to different number of receive elements at the receiver 130, e.g., multiple antennas, or a different sampling rate of the received signal. The difference in dimensionality of the version $\hat{z}$ of the transmitted signal input to the decoder compared to the quantized signal vector $\bar{z}$ output at the transmitter may also be due to some processing carried out at the receiver 130. For example, each detected signal may pre-processed and be provided to the decoder as a vector of likelihoods of assignment to the different symbol values for the communications channel output. That is, the decoder neural network 135 may receive, for each detected signal value, a vector of soft decisions on assignment to the different symbols, rather than a specific symbol value. In this respect, the noise-affected version $\hat{z}$ of the hard quantized signal values provided to the decoder may represent the received signal information differently to the representation provided to the modulator at the transmitter in the quantized signal vector $\bar{z}$. Nevertheless, regardless of its dimensionality and representation of the detected signal values, the noise-affected version $\hat{z}$ of the vector of hard quantized signal values transmitted over the noisy communication channel and provided to the decoder neural network 135 is simply one that is correlated with the input data from the information source 111. The decoder neural network 135 then learns to reconstruct the information source 111 from this noise-affected version $\hat{z}$ of the transmitted signals, regardless of its dimensionality and representation.

[00127] It should also be noted that, in accordance with the present disclosure, the soft and hard quantization steps can be carried out symbol-by-symbol or over blocks of symbols. That is, the input z to the soft quantization module may be plural latent vector elements in the signal space, which may be quantized together to provide plural soft quantized outputs. For example, z_1 and z_2 may be quantized together by the soft quantization module to provide at the output $\tilde{z}_1$ and $\tilde{z}_2$. Thus, it should be understood that the quantization provided in accordance with the present disclosure can be scalar quantization or vector quantization.

[00128] Throughout the description and claims of this specification, the words “comprise” and “contain” and variations of them mean “including but not limited to”, and they are not intended to (and do not) exclude other components, integers or steps. Throughout the description and claims of this specification, the singular encompasses the plural unless the context otherwise requires. In particular, where the indefinite article is used, the specification is to be understood as contemplating plurality as well as singularity, unless the context requires otherwise.

[00129] Features, integers, characteristics or groups described in conjunction with a particular aspect, embodiment or example of the invention are to be understood to be applicable to any other aspect, embodiment or example described herein unless incompatible therewith. All of the features disclosed in this specification (including any accompanying claims, abstract and drawings), and/or all of the steps of any method or process so disclosed, may be combined in

any combination, except combinations where at least some of such features and/or steps are mutually exclusive. The invention is not restricted to the details of any foregoing embodiments. The invention extends to any novel one, or any novel combination, of the features disclosed in this specification (including any accompanying claims, abstract and drawings), or to any novel one, or any novel combination, of the steps of any method or process so disclosed. In particular, any dependent claims may be combined with any of the independent claims and any of the other dependent claims.

CLAIMS

1. A communication system for conveying data from an information source across a communication channel using joint source and channel coding, comprising:

a transmitter including:

an encoder neural network having:

an encoder input layer having input nodes for receiving input data from an information source to be transmitted;

an encoder output layer having output nodes for outputting a vector of values in a signal space for modulating a carrier signal for transmission of a transformed version of the input data across a communication channel; and

one or more hidden layers having connecting nodes connecting the encoder input layer to the encoder output layer, the connecting nodes having weights to map the input data received at the encoder input layer to provide at the encoder output layer a transformed representation of the input data;

a soft quantization module configured to receive the unconstrained output signal space values from the encoder output layer and to operate a differentiable function over the signal space that transforms the unconstrained output signal space values to a smoothed approximation of a signal value in a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the transmitter over the communication channel; and

a hard quantization module configured to transform the approximate signal values output from the soft quantization module exactly to a signal value assigned to the symbol in the predetermined finite set S of the alphabet, the hard quantized signal values representing a transformed version of the input data and being passed as a vector from the hard quantization module to the transmitter to modulate a carrier signal for transmission over the communication channel;

and

a receiver including:

a decoder neural network having:

a decoder input layer having nodes corresponding to a channel output vector received at the receiver receiving the signal, the channel output vector corresponding to a noise-affected version of the vector of hard quantized signal values transmitted over the noisy communication channel;

a decoder output layer having output nodes for outputting a reconstructed representation of the input data provided from the information source to the encoder input layer of the encoder; and

one or more hidden layers having connecting nodes connecting the decoder input layer to the decoder output layer, the connecting nodes having weights to map the noise-affected version of the vector of hard quantized signal values received at the decoder input layer to provide at the decoder output layer a reconstructed representation of the input data;

the connecting node weights of the hidden layers of the encoder neural network and decoder neural network having been trained together to seek to minimise an objective function characterising a reconstruction error between the input-output pairs of training data, consisting of information source samples, which are passed to the encoder neural network, and the corresponding reconstructions at the decoder output layer of the decoder, using an appropriate optimization algorithm operating on the objective function..

2. A transmitter for conveying data from an information source across a communication channel using joint source and channel coding, the transmitter comprising:

an encoder neural network having:

an encoder input layer having input nodes for receiving input data from an information source to be transmitted;

an encoder output layer having output nodes for outputting a vector of unconstrained values in a signal space for modulating a carrier signal for transmission of a transformed version of the input data across a communication channel; and

one or more hidden layers having connecting nodes connecting the encoder input layer to the encoder output layer, the connecting nodes having weights to map the input data received at the encoder input layer to provide at the encoder output layer a transformed representation of the input data;

a soft quantization module configured to receive the unconstrained output signal space values from the encoder output layer and to operate a differentiable function over the signal space that transforms the unconstrained output signal space values to a smoothed approximation of a signal value in a predetermined finite set S of symbols of a predefined alphabet of carrier modulation signal values transmittable by the transmitter over the communication channel; and

a hard quantization module configured to transform the approximate signal values output from the soft quantization module exactly to a signal value assigned to the symbol in the predetermined finite set S of the alphabet, the hard quantized signal values representing a transformed version of the input data and being passed as a vector from the hard quantization

module to the transmitter to modulate a carrier signal for transmission over the communication channel;

the connecting node weights of the hidden layers of the encoder neural network having been trained together with the connecting node weights of the hidden layers of a decoder neural network for mapping a noise-affected version of the vector of hard quantized signal values received over the communication channel at the receiver to a reconstructed representation of the input data provided from the information source, the training seeking to minimise an objective function characterising a reconstruction error between the input-output pairs of training data, consisting of information source samples, which are passed to the encoder neural network, and the corresponding reconstructions at the decoder output layer of the decoder, using an appropriate optimization algorithm operating on the objective function.

3. A receiver for reconstructing data from an information source sent across a communication channel using joint source and channel coding comprising:

a decoder neural network having:

a decoder input layer having nodes corresponding to a channel output vector received at the receiver receiving a signal, the channel output vector corresponding to a noise-affected version of a vector of hard quantized signal values transmitted over the noisy communication channel by a transmitter including a corresponding encoder neural network, soft quantization module and hard quantization module, the transmitter being for mapping input data received from an information source to be transmitted to the vector of hard quantized signal values representing a transformed version of the input data;

a decoder output layer having output nodes for outputting a reconstructed representation of the input data provided from the information source to the encoder input layer of the encoder neural network; and

one or more hidden layers having connecting nodes connecting the decoder input layer to the decoder output layer, the connecting nodes having weights to map the noise-affected version of the vector of hard quantized signal values received at the decoder input layer to provide at the decoder output layer a reconstructed representation of the input data;

the connecting node weights of the hidden layers of the decoder neural network having been trained together with the connecting node weights of the hidden layers of the encoder neural network to seek to minimise an objective function characterising a reconstruction error between the input-output pairs of training data, consisting of information source samples, which are passed to the encoder neural network, and the corresponding reconstructions at the decoder output layer of the decoder, using an appropriate optimization algorithm operating on the objective function.

4. A communication system, transmitter, or receiver as claimed in any preceding claim, wherein the hard quantized signal values represent in-phase and quadrature components for modulation of the carrier signal for transmission over the communication channel.

5. A communication system, transmitter, or receiver as claimed in any preceding claim, wherein the predefined alphabet is a fixed, predefined constellation of symbols for digitally modulating the carrier signal to encode the input data for transmission over the communication channel.

6. A communication system, transmitter, or receiver as claimed in any preceding claim, wherein the predefined alphabet is selected to be one that compatible with the modulator of the transmitter and/or compatible with the demodulator of the receiver.

7. A communication system, transmitter, or receiver as claimed in any preceding claim, wherein the alphabet of symbols is mapped onto a constellation diagram for Phase Shift Keying or Quadrature Amplitude Modulation of the carrier signal, or wherein the alphabet of symbols is mapped onto the signal values for an existing inner channel code for the communications channel.

8. A communication system, transmitter, or receiver as claimed in any preceding claim, wherein the encoder neural network and decoder neural network have been trained together using a gradient descent algorithm operating on the objective function based on a differential of at least the decoder neural network, the soft quantization function and the encoder neural network

9. A communication system, transmitter, or receiver as claimed in any preceding claim, wherein the encoder neural network and decoder neural network have been trained together using training data in which a model of the communication channel is used to estimate channel noise and add it to the transmitted hard quantized signal values to generate a noise-affected version of the vector of hard quantized signal values in the input-output pairs of training data.

10. A communication system, transmitter, or receiver as claimed in claim 9, wherein during training the gradient descent algorithm has further operated on the objective function further based on a differential of the channel model.

11. A communication system, transmitter, or receiver as claimed in claim 9, wherein during training the gradient descent algorithm has operated on the objective function further based on a differentiable representation of the channel model generated using a generative adversarial network.

12. A communication system, transmitter, or receiver as claimed in any of claims 1 to 8, wherein the encoder neural network and decoder neural network have been trained together using training data in which input-output pairs of training data are transmitted over the communication channel by the transmitter, in order to add noise to the transmitted hard quantized signal values to generate the noise-affected version of the vector of hard quantized signal values in the input-output pairs of training data.

13. A communication system, transmitter, or receiver as claimed in any preceding claim, wherein the symbol alphabet set size S is a parameter of the soft and hard quantization modules, and wherein during training the gradient descent algorithm has further operated over a range of symbol alphabet set sizes S of different cardinality to learn the optimal modulation scheme for the information source.

14. A communication system, transmitter, or receiver as claimed in any preceding claim, wherein the objective function sought to be minimised by the training together of the encoder neural network and decoder neural network additionally characterises a relative entropy between the probability distribution of the symbols of the hard quantized signal values output from the hard quantization module and a uniform distribution across the symbols.

15. A communication system, transmitter, or receiver as claimed in claim 13, wherein the Kullback-Liebler divergence between the symbol distribution and the uniform distribution is used in the objective function to characterise the relative entropy for minimisation during training.

16. Method of training an encoder neural network and decoder neural network for use in a communication system as claimed in claim 1, for conveying data from an information source across a communication channel using joint source channel coding, the method comprising:

for input-output pairs of a set of training data from the information source passed to the encoder neural network, determining an objective function characterising a reconstruction error between input-output pairs of training data from the information source passed to the encoder neural network and the representation of the input data reconstructed at the decoder output layer;

using an appropriate optimisation algorithm operating on the objective function, updating the connecting node weights of the hidden layers of the encoder neural network and decoder neural network to seek to minimise the objective function.

Intellectual Property Office

Application No:	GB2107733.4	**Examiner:**	Contract Unit Examiner
Claims searched:	1-16	**Date of search:**	9 March 2022

Patents Act 1977: Search Report under Section 17

Documents considered to be relevant:

Category	Relevant to claims	Identity of document and passage or figure of particular relevance
X	3-16	IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, vol 1, 2020, LEINONEN MARKUS ET AL, "Low-Complexity Vector Quantized Compressed Sensing via Deep Neural Networks", pages 1278-1294, DOI:10.1109/OJCOMS.2020.3020131 See Section I-IV
X	3-16	WO2020/035684 A1 (IMPERIAL COLLEGE SCI TECH & MEDICINE) paragraphs [0144] - [0181]; figures 1-4
X	3-16	US 2020/0304802 A1 (HABIBIAN AMIRHOSSEIN ET AL) paragraphs [0069] - [0110]; figures 4A, 4B, 5, 6
A	-	ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 2018, JUN HAN ET AL, "Deep Probabilistic Video Compression", available from https://arxiv.org/pdf/1810.02845v1.pdf See Section 3, 4

Categories:

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.

Field of Search:

Search of GB, EP, WO & US patent documents classified in the following areas of the UKC^{X} :

Worldwide search of patent documents classified in the following areas of the IPC

G06N; H04N

The following online and other databases have been used in the preparation of this search report

Intellectual Property Office

International Classification:

Subclass	Subgroup	Valid From
G06N	0003/04	01/01/2006
G06N	0003/08	01/01/2006
H04N	0019/124	01/01/2014