CLAIMS

1. A method for entropy encoding of a latent tensor (4020), comprising:

separating (700, 701) the latent tensor (4020) into a plurality of segments (820, 920) in the spatial dimensions and in the channel dimension, each segment including at least one latent tensor element;

processing an arrangement (830, 930, 940) of the plurality of segments by one or more layers of a neural network, including at least one attention layer (732); and

obtaining a probability model (770) for the entropy encoding of a current element of the latent tensor (4020) based on the processed plurality of segments (820, 920).

2. The method according to claim 1, wherein the processing of the arrangement of the plurality of segments further comprises:

obtaining two or more groups (1520, 1521, 1522) of segments; and

processing the segments within a group out of the two or more groups (1520, 1521, 1522) independently by the one or more layers of the neural network.

3. The method according to claim 2, wherein for each group out of the two or more groups (1520, 1521, 1522), the segments in said group have a same respective channel segment index, said channel segment index indicating the segments the channel dimension.

4. The method according to any of the claims 2 to 3, wherein the segments having a same channel segment index are grouped into either two groups (1520, 1521) or four groups (1620, 1621, 1622, 1632), wherein the segments having a same channel segment index are grouped according to a checkerboard pattern into either two groups (1520, 1521) or four groups (1620, 1621, 1622, 1623) in the spatial dimensions.

5. The method according to claim 3 or 4, wherein the processing of the arrangement of the plurality of segments further comprises:

i) processing the segments in a first group (1520, 1620) in parallel, followed by

processing the segments in a second group (1521, 1621) in parallel, wherein the segments in the second group have a same index in the channel dimension as the segments in the first group (1520, 1620), or

ii) processing the segments in a first group (1530, 1630) in parallel, followed by

processing the segments in a second group (1531,1631) in parallel, wherein the segments in the second group have a same spatial coordinate as the corresponding segments in the first group (1530, 1630).

6. The method according to any of the claims 2 to 5, further comprising

generating a bitstream (371); and

including an indication into said bitstream (371) indicating whether or not two or more groups of segments are obtained.

7. A method for entropy decoding of a latent tensor (4020), comprising:

initializing the latent tensor (4020) with zeroes;

separating (700, 701) the latent tensor (4020) into a plurality of segments (820, 920) in the spatial dimensions and in the channel dimension, each segment including at least one latent tensor element;

processing an arrangement (830, 930, 940) of the plurality of segments by one or more layers of a neural network, including at least one attention layer (732); and

obtaining a probability model (770) for the entropy decoding of a current element of the latent tensor (4020) based on the processed plurality of segments (820, 920).

8. The method according to claim 7, wherein the processing of the arrangement of the plurality of segments further comprises:

obtaining two or more groups (1520, 1521, 1522) of segments; and

processing the segments within a group out of the two or more groups (1520, 1521, 1522) independently by the one or more layers of the neural network.

9. The method according to claim 8, wherein for each group out of the two or more groups (1520, 1521, 1522), the segments in said group have a same respective channel segment index, said channel segment index indicating the segments the channel dimension.

10. The method according to claim 8 or 9, wherein the segments having a same channel segment index are grouped into either two groups (1520, 1521) or four groups (1620, 1621, 1622, 1632).

11. The method according to claim 10, wherein the segments having a same channel segment index are grouped according to a checkerboard pattern into either two groups (1520, 1521) or four groups (1620, 1621, 1622, 1632) in the spatial dimensions.

12. The method according to any of the claims 9 to 11, wherein the processing of the arrangement of the plurality of segments further comprises:

i) processing the segments in a first group (1520, 1620) in parallel, followed by

processing the segments in a second group (1521, 1621) in parallel, wherein the segments in the second group have a same index in the cannel dimension as the segments in the first group (1520, 1620), or

ii) processing the segments in a first group (1530, 1630) in parallel, followed by

processing the segments in a second group (1531, 1631) in parallel, wherein the segments in the second group have a same spatial coordinate as the corresponding segments in the first group (1530, 1630).

13. The method according to any of the claims 8 to 12, further comprising

receiving a bitstream (371); and

obtaining an indication from said bitstream (371) indicating whether or not two or more groups of segments are obtained.

14. The method according to claims 7 to 13, wherein the processing of the arrangement further of the plurality of segments comprises

separating the segments into a plurality of patches (1920, 1921, 1922, 1923), each patch including two or more segments; and

processing each patch independently by the one or more layers of the neural network.

15. The method according to claim 14, wherein each patch includes a K x M grid of spatially neighboring segments and L neighboring segments in the channel dimension, with L, K and M being positive integers, at least one of K and M being larger than one.

16. The method according to claim 14 or 15, wherein the patches out of the plurality of patches (1930, 1931, 1932, 1933) are overlapping in the spatial dimensions and/or in the channel dimension.

17. The method according to any of the claims 14 to 16, wherein the processing of a patch out of the plurality of patches (1920, 1921, 1922, 1923) by the one or more layers of the neural network further comprises

obtaining two or more groups (1520, 1521, 1522) of segments within said patch; and

processing the segments within a group out of the two or more groups (1520, 1521, 1522) independently by the one or more layers of the neural network.

18. The method according to any of the claims 7 to 17, wherein the processing of the arrangement comprises arranging the segments in a predefined order (930), wherein segments with a same spatial coordinate are grouped together, or

wherein the processing of the arrangement comprises arranging the segments, wherein segments, which have different spatial coordinates, are arranged consecutively in a predefined order (940).

19. The method according to any of claims 7 to 18, wherein the processing by the neural network comprises applying a first neural subnetwork (720) to extract features of the plurality of segments (820, 920), and providing an output of the first neural subnetwork (720) as an input to a subsequent layer within the neural network.

20. The method according any of claims 7 to 19, wherein the processing by the neural network further comprises:

providing positional information (721) of the plurality of segments (820, 920) as an input to the at least one attention layer (732).

21. The method according to any of the claims 7 to 20 wherein the processing an arrangement (830, 930, 940) of the plurality of segments includes selecting a subset of segments from said plurality of segments; and said subset is provided as an input to a subsequent layer within the neural network.

22. The method according to any of the claims 7 to 21, wherein the neural network includes a second neural subnetwork (735), the second neural subnetwork (735) processing an output of the attention layer (732).

23. The method according to any of the claims 7 to 22, wherein the at least one attention layer (732) in the neural network is a multi-head attention layer.

24. The method according to any of the claims 7 to 23, wherein the at least one attention layer (732) in the neural network is included in a transformer subnetwork, and

padding (710, 711) the beginning of the arrangement of the plurality of segments with a zero segment before processing by the neural network.

25. The method according to any of the claims 7 to 24, further comprising:

entropy decoding the current element into a first bitstream (371) using the obtained probability model (770).

26. A method for decoding image data comprising:

entropy decoding a latent tensor (4020) from a bitstream (371) according to claims 7 to 25; and

obtaining the image data by processing the latent tensor (4020) with an autodecoding convolutional neural network.

27. A computer program stored on a non-transitory medium and including code instructions, which, when executed on one or more processors, cause the one or more processors to execute steps of the method according to any of claims 1 to 26.

28. An apparatus for entropy encoding of a latent tensor, comprising:

processing circuitry configured to:

separate (700, 701) the latent tensor (4020) into a plurality of segments (820, 920) in the spatial dimensions and in the channel dimension, each segment including at least one latent tensor element;

process an arrangement of the plurality of segments (830, 930, 940) by one or more layers of a neural network, including at least one attention layer (732); and

obtain a probability model (770) for the entropy encoding of a current element of the latent tensor (4020) based on the processed plurality of segments (820, 920).

29. An apparatus for entropy decoding of a latent tensor, comprising:

processing circuitry configured to:

initialize the latent tensor (4020) with zeroes;

separate (700, 701) the latent tensor (4020) into a plurality of segments (820, 920) in the spatial dimensions and in the channel dimension, each segment including at least one latent tensor element;

process an arrangement (830, 930, 940) of the plurality of segments by one or more layers of a neural network, including at least one attention layer (732); and

obtain a probability model (770) for the entropy decoding of a current element of the latent tensor (4020) based on the processed plurality of segments (820, 920).