US 20250259421A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2025/0259421 A1**

**Pototzky** (43) **Pub. Date: Aug. 14, 2025**

(54) **FASTER CONVERGING PRE-TRAINING FOR MACHINE LEARNING MODELS**

(71) Applicant: **Robert Bosch GmbH**, Stuttgart (DE)

(72) Inventor: **Daniel Pototzky**, Hildesheim (DE)

(21) Appl. No.: **18/850,095**

(22) PCT Filed: **May 3, 2023**

(86) PCT No.: **PCT/EP2023/061645**

§ 371 (c)(1),
(2) Date: **Sep. 24, 2024**

(30) **Foreign Application Priority Data**

May 6, 2022 (DE) ..... 10 2022 204 492.4

**Publication Classification**

(51) **Int. Cl.**

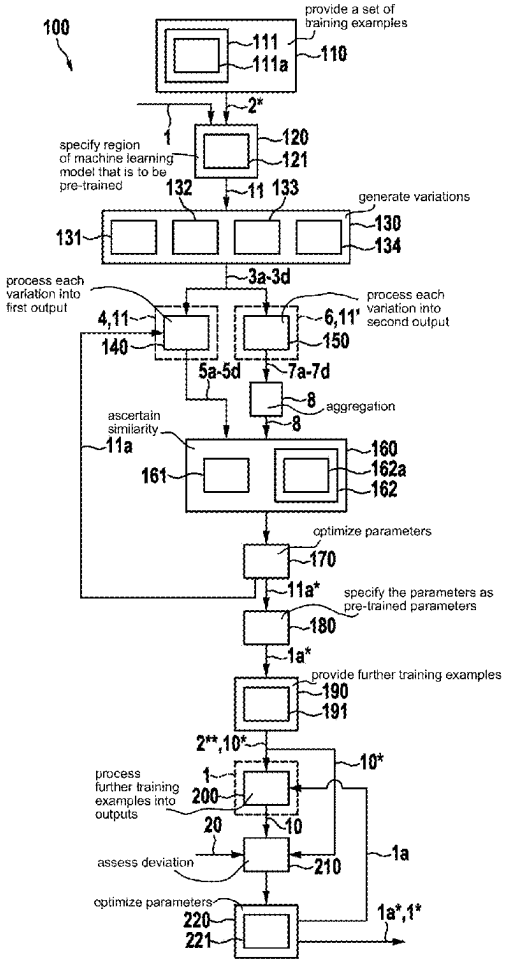
| | |
|---|---|
| ***G06V 10/774*** | (2022.01) |
| ***G06V 10/72*** | (2022.01) |
| ***G06V 10/764*** | (2022.01) |
| ***G06V 10/77*** | (2022.01) |
| ***G06V 10/776*** | (2022.01) |
| ***G06V 20/58*** | (2022.01) |

(52) **U.S. Cl.**

CPC ..... ***G06V 10/774*** (2022.01); ***G06V 10/72*** (2022.01); ***G06V 10/764*** (2022.01); ***G06V 10/7715*** (2022.01); ***G06V 10/776*** (2022.01); ***G06V 20/58*** (2022.01)

(57) **ABSTRACT**

A method for unsupervised pre-training of a machine learning model. The method includes providing a set of training examples for inputs of the machine learning model; specifying a region of the machine learning model to be pre-trained; generating variations from each training example; processing each variation into a first output in a first processing branch, which includes at least one first instance of the region to be pre-trained; processing each variation into a second output in a second processing branch which includes at least one second instance of the region to be pre-trained; for each variation, ascertaining the similarity of the first output generated from this variation to an aggregation of the second outputs generated from all the other variations of the same training example; optimizing parameters that characterize behavior of the first instance of the region to be pre-trained, with the goal of maximizing the similarity thus ascertained.



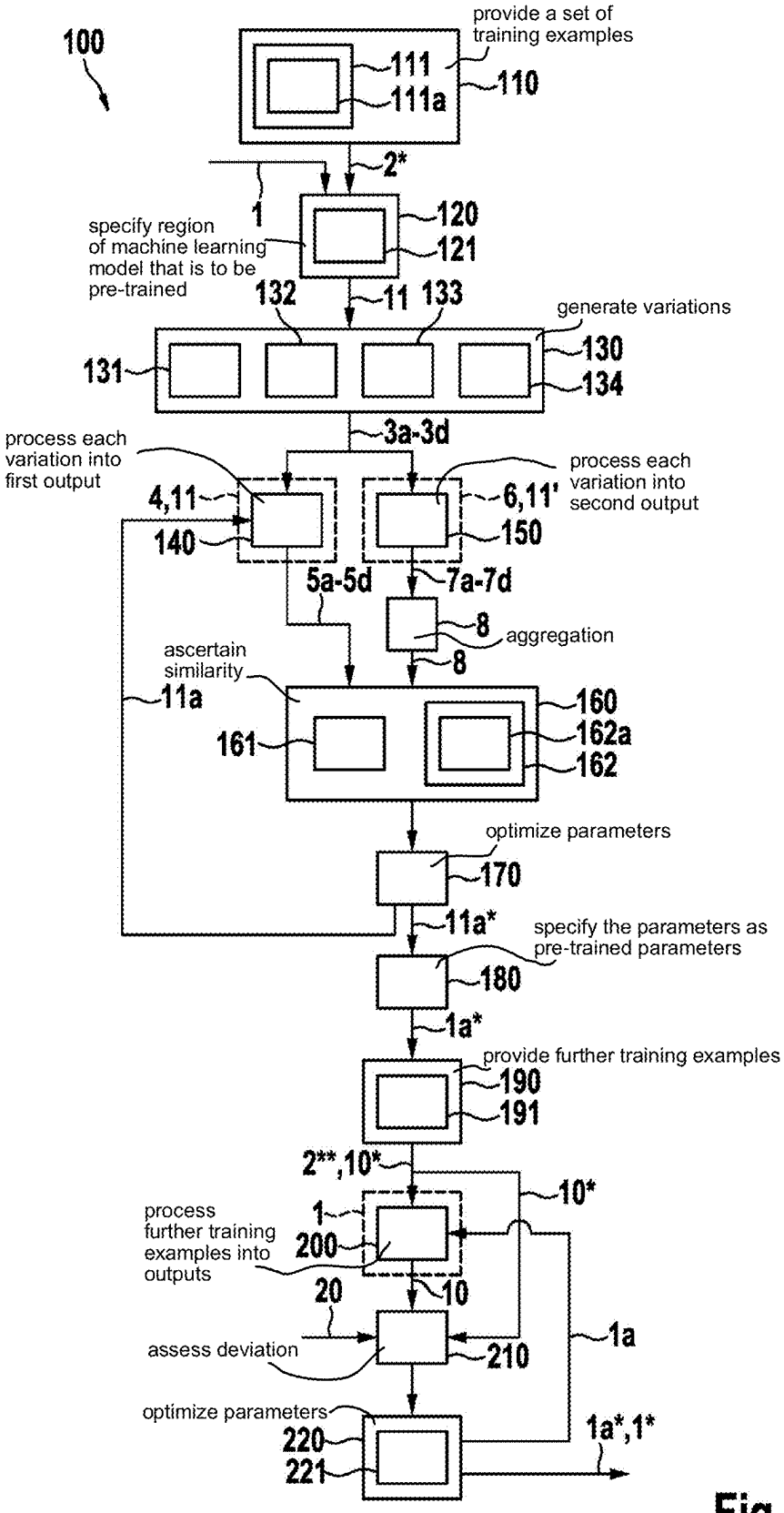

100
provide a set of training examples
111
111a
110
2*
1
specify region of machine learning model that is to be pre-trained
120
121
132
11
133
generate variations
131
130
134
3a-3d
process each variation into first output
process each variation into second output
4,11
6,11'
140
150
5a-5d
7a-7d
8
aggregation
8
ascertain similarity
11a
160
161
162a
162
optimize parameters
170
11a*
specify the parameters as pre-trained parameters
180
1a*
provide further training examples
190
191
2**,10*
10*
process further training examples into outputs
1
200
20
10
assess deviation
210
1a
optimize parameters
220
221
1a*,1*

Fig. 1

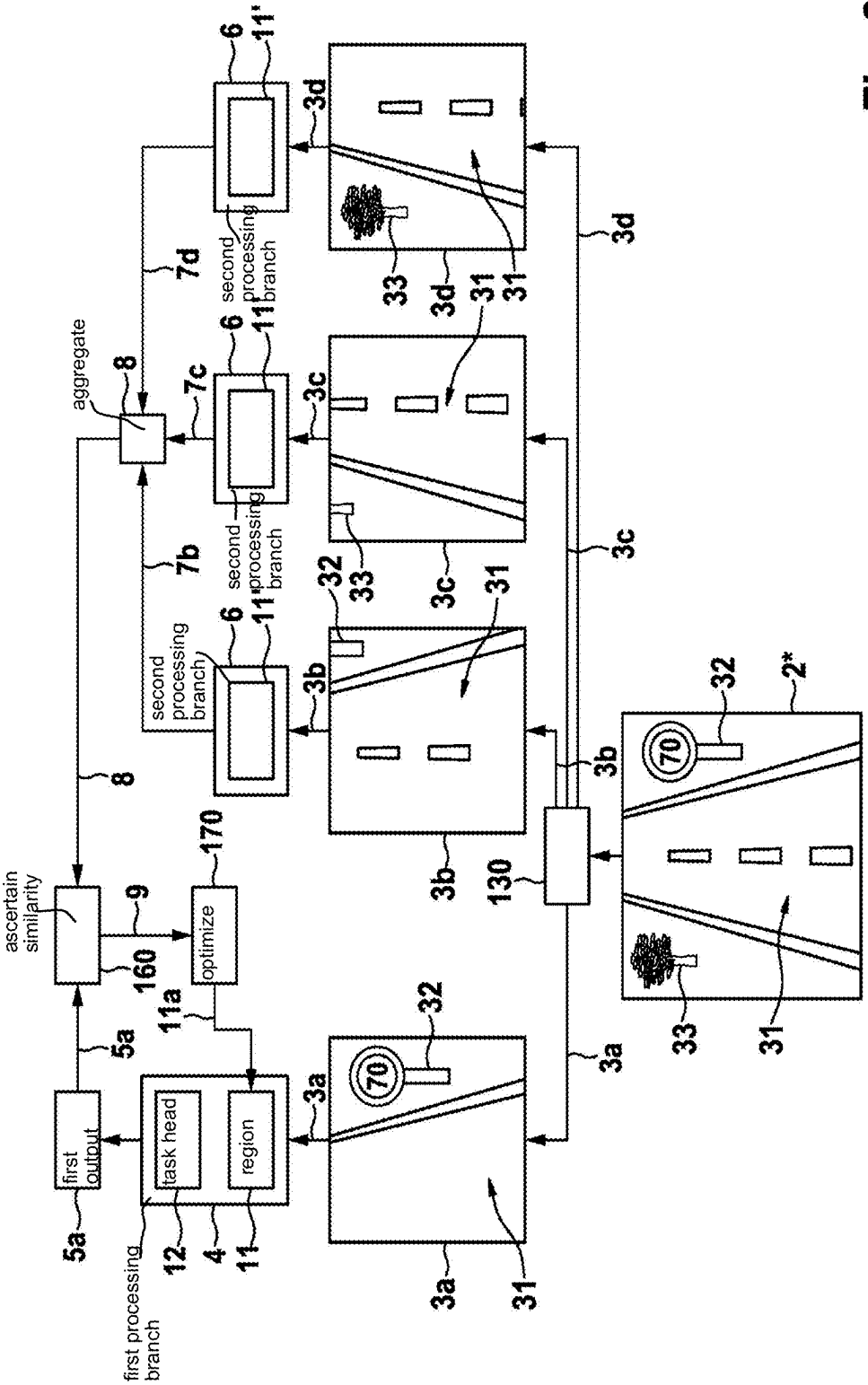

first processing branch
second processing branch
second processing branch
second processing branch
second processing branch
first output
task head
region
ascertain similarity
optimize
aggregate
5a
12
4
11
160
9
170
11a
8
7b
7c
7d
6
11'
11
3a
3b
3c
3d
31
32
33
70
130
2*

Fig. 2

# FASTER CONVERGING PRE-TRAINING FOR MACHINE LEARNING MODELS

## FIELD

**[0001]** The present invention relates to the unsupervised pre-training of machine learning models, which can then be further trained, for example in a supervised manner, for a given task.

## BACKGROUND INFORMATION

**[0002]** Training a machine learning model on a particular task, starting from an empty or randomly initialized starting state, requires a large number of training examples and a lot of computing time. A significant portion of this effort can be saved by using a generically pre-trained machine learning model and then further training it specifically for the desired task. Often, the same generically pre-trained model can be used as a starting point for the training for many tasks. For example, an image classifier can be generically pre-trained to recognize particular basic features in the images. In the specific training, the image classifier can then, for example, learn to recognize particular objects by using these basic features.

**[0003]** Generic pre-training promises a large cost advantage, especially if this pre-training takes place in an unsupervised manner, i.e., on the basis of training examples that are not labeled with target outputs. Labeling training examples is very expensive since it typically requires manual work.

## SUMMARY

**[0004]** The present invention provides a method for unsupervised pre-training of a machine learning model.

**[0005]** A machine learning model in particular refers to a model that embodies a function that is parameterized with adjustable parameters and has great power to generalize. When training a machine learning model, the parameters can in particular be adjusted in such a way that, when training examples are entered into the machine learning model, the associated target outputs are reproduced as well as possible. The machine learning may in particular include an artificial neural network (ANlN) and/or be an ANN.

**[0006]** According to an example embodiment of the present invention, within the framework of the method, a set of training examples is provided for inputs of the machine learning model. These training examples do not need to be labeled with target outputs.

**[0007]** A region of the machine learning model that is to be pre-trained is also specified. This region may in particular be selected such that it can be used in the pre-trained state for multiple different tasks. For example, many classifiers comprise a region that analyzes the input of the machine learning model for features and a task head that uses the result of this analysis to ascertain classification scores with respect to one or more classes of a given classification. The task head is then specific to the specific classification task, while the analyzed features can also be used for other tasks. It then makes sense not to include the task head in the pre-training.

**[0008]** Thus, a region of the machine learning model that is designed to extract features from the input of the machine learning model is advantageously selected as a region to be pre-trained. In this context, it is once again advantageous that the training examples do not have to be labeled, because the available labels generally relate to the final output of the machine learning model with respect to the given task. However, no labels are available for intermediate results, such as feature maps of a feature extractor. It is also not possible, for example, to infer target outputs regarding the feature maps from target outputs regarding classification scores to be provided by the machine learning model, because the task head maps many different feature maps, which were generated, for example, for different images of one and the same object from different perspectives, to one and the same vector of classification scores, i.e., for example, a "one-hot vector" with a score of 1 for the class of the object.

**[0009]** Variations are generated from each training example. Any data augmentation method can be used for this purpose.

**[0010]** Each variation is processed into a first output in a first processing branch, which comprises at least one first instance of the region to be pre-trained. However, each variation is also processed into a second output in a second processing branch, which comprises at least one second instance of the region to be pre-trained. In this context, the term "instances" is in particular to be understood to mean that the processing in both instances takes place independently of one another. In particular, it should not affect the processing in the second instance if parameters (for example, weights) of the first instance are changed.

**[0011]** For each variation, the similarity of the first output generated from this variation to an aggregation of the second outputs generated from all other variations of the same training example is ascertained. Parameters that characterize the behavior of the first instance of the region to be pre-trained are optimized with the goal of maximizing the similarity thus ascertained. The thus obtained parameters are specified as pre-trained parameters of the machine learning model.

**[0012]** It was found that especially aggregating the second outputs generated from the respective other variations of the same training example makes the pre-training significantly more stable. These second outputs provide, in a sense, the goal at which the optimization in the first instance of the region to be pre-trained is aimed. When individual comparisons are made between first outputs and second outputs, the second output and thus the goal changes with each new variation of the training example. However, training in which the goal constantly changes takes significantly longer and also requires significantly more resources, for example multiple GPUs, the cooperation of which must also be synchronized. This is somewhat analogous to the scenario where, when constructing a building, it is detrimental for adhering to the schedule and budget if the client puts forward new requests for changes to the planning every day while the work is already underway. However, if a plan is consistently followed, it can work.

**[0013]** According to the related art, which was based on individual comparisons, large batches of training examples were needed and training had to take place over many epochs until the pre-training converged to a stable result. This caused a certain averaging effect over the different optimization goals that respectively resulted from the different variations. However, it is much faster and also much easier to accomplish to incorporate the aggregation effect

into the specification of the optimization goal and then to consistently pursue this one optimization goal.

**[0014]** This is also evident when analyzing the statistics. For example, let $T_1$ be a second output generated from a particular variation of a training example. If the variations are generated randomly, it can be assumed that $T_1$ is a sample from a normal distribution with a mean value p and a standard deviation $\sigma$:

$$T_1 \sim N(\mu, \sigma^2 I),$$

where I is the identity matrix. However, this sample as such can still deviate significantly from what would most often be expected in the normal distribution. However, if K samples are taken and averaged, the variance $\sigma^2$ is reduced by the factor K. The effect is the same as if a single sample $\overline{(T_K)}$ were taken from a distribution with the reduced variance:

$$\overline{T_K} \sim N\left(\mu, \frac{\sigma^2 I}{K}\right).$$

**[0015]** Even if the simplifying assumption that the distribution is a normal Gaussian distribution is not valid, $\overline{T_K} \to \mu$ is still true for $K \to \infty$ as long as the samples are independently and identically distributed.

**[0016]** For generating variations of a training example, any data augmentation method the result of which can still be unambiguously traced back to exactly this training example is suitable. For example, a proper subset of the data of the training example can be randomly selected, such as an image section from a training image, a subarea of a point cloud, or a temporal section from a time series. It is also possible, for example, to impress noise sampled from a random distribution on the data of the training example. Even with this noise, the original training example can still be recognized. The same applies if a particular portion of the data of the training example is removed or made unrecognizable in whole or in part, such as by blurring or blacking out a subregion in an image. As an alternative to or in combination with the options mentioned so far, any other transformation that does not change the semantic content of the data can be applied to the data of the training example.

**[0017]** According to an example embodiment of the present invention, in particular, the aggregation may, for example, comprise ascertaining a mean value, a medoid, or an element-wise maximum. Other aggregations are also suitable insofar as they smooth out differences between the second outputs and, in particular, suppress the influence of outliers.

**[0018]** According to an example embodiment of the present invention, the similarity can, for example, in particular be ascertained by means of a distance measure. For example, if the region of the machine learning model that is to be pre-trained extracts features from the input of the machine learning model, such a distance measure is particularly easy to interpret in the space of the ascertained features. For example, a cosine distance can in particular be chosen as a distance measure.

**[0019]** Starting from a first output $P_{K+1}$ for a training example and K variations $z_1, \ldots, z_K$, a cost function (loss function)

$$L(z_1, \ldots, z_K, p_{K+1}) = -\frac{p_{K+1}}{\|p_{K+1}\|_2} \cdot \frac{\frac{1}{K}\sum_{i=1}^{K} z_i}{\left\|\frac{1}{K}\sum_{i=1}^{K} z_i\right\|_2} = -\frac{p_{K+1}}{\|p_{K+1}\|_2} \cdot \frac{\overline{T_K}}{\|\overline{T_K}\|_2}$$

can thus, for example, be used for the pre-training, where $\overline{(T_K)}$ is a sample from a normal distribution of which the variance has been reduced by the factor K through aggregation. If each variation is processed once into the first output and the respective other variations specify the optimization goal, the overall result is the cost function

$$\mathcal{L}_{total} = \sum_{i=1}^{K+1} \frac{1}{K+1} \mathcal{L}(\{z_j \mid j \neq i \wedge 1 \leq j \leq K+1\}, p_i)$$

for the pre-training.

**[0020]** The training examples may, for example, in particular comprise images or point clouds that were recorded by measurement observation of a scene. Images can be still images or moving images, which were, for example, recorded with one or more cameras for visible light or other parts of the electromagnetic spectrum (such as infrared). Point clouds can be recorded by means of radar sensors or lidar sensors, for example. Especially these data types have a particularly high dimensionality, so that the stability of the training goal that was gained by aggregating the second outputs is particularly important.

**[0021]** The scene can, for example, in particular be a traffic situation that can be observed from a vehicle. It is particularly complex to obtain labeled training examples for traffic situations. Pre-training on unlabeled data is therefore particularly advantageous. Furthermore, especially if evaluating traffic situations, for example for the purposes of at least partially automated driving, there are a large number of tasks that can all utilize the pre-training.

**[0022]** The ultimate goal of the pre-training is to create a better basis for training the machine learning model on a specific given task. After the pre-training, it is therefore advantageous to provide further training examples for inputs of the machine learning model. These further training examples are labeled with target outputs with respect to a given task. They are processed into outputs by the machine learning model. Any deviation of these outputs from the target outputs is assessed by means of a given cost function. Parameters that characterize the behavior of the machine learning model are optimized with the goal that the assessment by the cost function is expected to improve during the further processing of labeled training examples.

**[0023]** In this context, the pre-training described above has the effect that the further training with regard to the specific task requires fewer labeled training examples and also converges faster. As a result, the total training effort, which includes the unsupervised generic pre-training and the supervised task-specific training, can be significantly reduced. Taking into account that the same generic pre-training can form the basis for many task-specific training processes, the savings become even greater.

**[0024]** The machine learning model can, for example, in particular be designed to ascertain a classification of the images or point clouds, which can, for example, in particular depend on the pixel values and/or values of measured

variables in these images or point clouds. The classification can, for example, in particular also comprise a semantic segmentation as a region-wise or pixel-wise classification, or a detection of whether an object is present or not. As explained above, a machine learning model for such a task consists largely of regions that extract features from the inputs. Only a small part of the model is apportioned to the task head, which is specific to the specific classification task or segmentation task. A large part of the machine learning model can thus be pre-trained in an unsupervised manner, and the portion of the training that still has to be completed with labeled training examples becomes smaller.

**[0025]** During further training with the labeled training examples, the already pre-trained parameters can be further optimized. With regard to the pre-training, it is not asserted that feature extraction can, for example, lead to an optimum that is also an optimum with respect to every possible downstream task. In fact, different features may also become particularly relevant with respect to different tasks, and the ultimately achieved task accuracy can be improved by focusing the feature extraction precisely on these features, possibly at the expense of other features. This can, for example, be illustrated on the basis of the example of image processing: The contrast of particular desired features can possibly only be increased at the cost of bringing other, less important features into saturation.

**[0026]** However, in a further advantageous example embodiment of the present invention, it may again be advantageous to retain the pre-trained parameters during the training with the further, labeled training examples. This can be useful, for example, if the pre-trained region of the machine learning model has been officially certified in precisely this state. Such certifications may, for example, be required when using machine learning models to control vehicles or other safety-relevant systems.

**[0027]** In a further advantageous example embodiment of the present invention, the further, labeled training examples belong to a different distribution or domain than the training examples used for the pre-training. The method then makes use of the fact that the pre-training on one distribution or domain generalizes well to other distributions or domains.

**[0028]** The use of the terms "distribution" and "domain" is not to be understood as a restriction to the effect that a distribution or domain must be specified, from which the training examples are then sampled. In fact, a given set of training examples may also define a distribution or domain. For example, if all or almost all of the training samples are consistent with a normal distribution with a particular mean value and a particular standard deviation, or with a different random distribution of a specific type with specific parameters, the set of training samples defines this random distribution by allowing inference about its type and parameters. Different domains can, for example, in particular represent different circumstances under which measurement data were, for example, obtained as training examples, such as different seasons, times of day, weather conditions, or imaging modalities with which images were recorded as training examples.

**[0029]** The method according to the present invention can in particular be wholly or partially computer-implemented. For this reason, the present invention also relates to a computer program comprising machine-readable instructions which, when executed on one or more computers, cause said computer(s) to carry out the method described above. In this sense, control devices for vehicles and embedded systems for technical devices, which are also capable of executing machine-readable instructions, are also to be regarded as computers.

**[0030]** The present invention also relates to a machine-readable data carrier and/or to a download product comprising the computer program of the present invention. A download product is a digital product that can be transmitted via a data network, i.e., can be downloaded by a user of the data network, and can, for example, be offered for immediate download in an online shop.

**[0031]** Furthermore, a computer can be equipped with the computer program of the present invention, with the machine-readable data carrier, or with the download product.

**[0032]** Further measures improving the present invention are explained in more detail below, together with the description of the preferred exemplary embodiments of the present invention, with reference to figures.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0033]** FIG. **1** shows an exemplary embodiment of the method **100** for unsupervised pre-training of a machine learning model **1**, according to the present invention.

**[0034]** FIG. **2** shows an exemplary application of the method **100** on the basis of a training example **2*** of a traffic situation, according to the present invention.

## DETAILED DESCRIPTION OF EXAMPLE EMBODIMENTS

**[0035]** FIG. **1** is a schematic flow chart of an exemplary embodiment of the method **100** for unsupervised pre-training of a machine learning model **1**.

**[0036]** In step **110**, a set of training examples **2*** are provided for inputs **2** of the machine learning model **1**.

**[0037]** According to block **111**, images or point clouds recorded by measurement observation of a scene can be selected as training examples **2***. According to block **111***a*, a traffic situation that can be observed from a vehicle can be selected as a scene.

**[0038]** In step **120**, a region **11** of the machine learning model **1** that is to be pre-trained is specified.

**[0039]** According to block **121**, a region of the machine learning model **1** that is designed to extract features from the input **2** of the machine learning model **1** can be selected as a region **11** to be pre-trained.

**[0040]** In step **130**, variations **3***a*-**3***d* are generated from each training example.

**[0041]** According to block **131**, a proper subset of the data of the training example **2*** can be randomly selected.

**[0042]** According to block **132**, noise sampled from a random distribution can be impressed on the data of the training example **2***.

**[0043]** According to block **133**, a portion of the data of the training example **2*** can be removed or made unrecognizable in whole or in part.

**[0044]** According to block **134**, a transformation, that does not change the semantic content of the data, can be applied to the data of the training example **2***.

**[0045]** In step **140**, each variation **3***a*-**3***d* is processed into a first output **5***a*-**5***d* in a first processing branch **4**, which comprises at least one first instance of the region **11** to be pre-trained.

**[0046]** In step **150**, each variation **3***a*-**3***d* is processed into a second output **7***a*-**7***d* in a second processing branch **6**, which comprises at least one second instance **11'** of the region **11** to be pre-trained.

**[0047]** In step **160**, for each variation **3***a*-**3***d*, the similarity **9** of the first output **5***a*-**5***d* generated from this variation **3***a*-**3***d* to an aggregation **8** of the second outputs **7***a*-**7***d* generated from all the other variations **3***a*-**3***d* of the same training example **2*** is ascertained.

**[0048]** According to block **161**, the aggregation **8** can comprise ascertaining a mean value, a medoid, or an element-wise maximum.

**[0049]** According to block **162**, the similarity **9** can be ascertained by means of a distance measure. According to block **162***a*, a cosine distance can be selected as a distance measure.

**[0050]** In step **170**, parameters **11***a* that characterize the behavior of the first instance of the region **11** to be pre-trained are optimized with the goal of maximizing the similarity **9**. The fully optimized state of the parameters **11***a* is denoted by reference sign **11***a**.

**[0051]** In step **180**, the parameters **11***a** are specified as pre-trained parameters **1***a* #of the machine learning model **1**.

**[0052]** In step **190**, further training examples **2**** are provided for inputs **2** of the machine learning model **1**. These further training examples are labeled with target outputs **10*** with respect to a given task.

**[0053]** According to block **191**, the further training examples **2**** may belong to a different distribution or domain than the training examples **2*** used for the pre-training.

**[0054]** In step **200**, the further training examples **2**** are processed into outputs **10** by the machine learning model **1**.

**[0055]** In step **210**, a deviation of these outputs **10** from the target outputs **10*** is assessed by means of a given cost function **20**.

**[0056]** In step **220**, parameters **1***a* that characterize the behavior of the machine learning model **1** are optimized with the goal that the assessment **20***a* by the cost function **20** is expected to improve during the further processing of labeled training examples **2****.

**[0057]** The fully trained state of the parameters **1***a* is denoted by reference sign **1***a**. The complete training also specifies the fully trained machine learning model **1*** as a whole.

**[0058]** According to block **221**, the pre-trained parameters **1***a* #can be retained during the training with the further training examples **2****.

**[0059]** FIG. **2** shows, by way of example, the application of the method **100** on the basis of a training example **2***, which is an image of a traffic situation. The traffic situation contains a road **31** as well as a tree **33** at the left-hand edge of the road **31** and a traffic sign **32** at the right-hand edge of the road.

**[0060]** In step **130** of the method **100**, variations **3***a*-**3***d* are generated by respectively selecting sections from the training example **2***. The variation **3***a* shows a part of the right-hand edge of the road **31** and the traffic sign **32**. The variation **3***b* shows another part of the right-hand edge of the road **31** and a part of the traffic sign **32**. The variation **3***c* shows a part of the left-hand edge of the road **31** and a part of the tree **33**. The variation **3***d* shows another part of the left-hand edge of the road **31** and the tree **33**.

**[0061]** In the example shown in FIG. **2**, the variation **3***a* is processed into the output **5***a* in the first processing branch **4**. The first processing branch **4** contains the region **11** of the machine learning model **1** to be pre-trained as well as the task head **12** of the machine learning model **1**. The variations **3***b*-**3***d* are processed into outputs **7***b*-**7***d* in a second processing branch **6**. The second processing branch **6** contains another instance **11'** of the region **11** of the machine learning model **1** that is to be pre-trained.

**[0062]** The outputs **7***b*-**7***d* are combined in an aggregation **8**. The output **5***a* from the first processing branch **4** is compared with this aggregation **8** in step **160**. On the basis of the ascertained similarity **9**, the parameters **11***a* of the region **11** to be pre-trained are optimized.

**1**-**15**. (canceled)

**16**. A method for unsupervised pre-training of a machine learning model, comprising the following steps:

providing a set of training examples for inputs of the machine learning model;

specifying a region of the machine learning model that is to be pre-trained;

generating variations from each of the training examples;

processing each of the variations into a first output in a first processing branch, the first processing branch including at least one first instance of the region to be pre-trained;

processing each of the variations into a second output in a second processing branch, the second processing branch including at least one second instance of the region to be pre-trained;

for each of the variations, ascertaining a similarity of the first output generated from the variation to an aggregation of the second outputs generated from all of the other variations of the same training example;

optimizing parameters that characterize a behavior of the first instance of the region to be pre-trained, with a goal of maximizing the ascertained similarity;

specifying the optimized parameters as pre-trained parameters of the machine learning model.

**17**. The method according to claim **16**, wherein the generating of the variations of each training example includes:

selecting a proper subset of data of the training example randomly; and/or

impressing noise sampled from a random distribution on the data of the training example; and/or

removing or making unrecognizable in whole or in part a portion of the data of the training example; and/or

applying, to the data of the training example, a transformation that does not change a semantic content of the data of the training example.

**18**. The method according to claim **16**, wherein the region to be pre-trained is a region of the machine learning model that is configured to extract features from the input of the machine learning model.

**19**. The method according to claim **16**, wherein the aggregation includes ascertaining a mean value, or a medoid, or an element-wise maximum.

**20**. The method according to claim **16**, wherein the similarity is ascertained using a distance measure.

**21**. The method according to claim **20**, wherein a cosine distance is the distance measure.

**22**. The method according to claim **16**, wherein images or point clouds recorded by measurement observation of a scene are the training examples.

**23**. The method according to claim **22**, wherein a traffic situation that can be observed from a vehicle the scene.

**24**. The method according to claim **16**, wherein:

further training examples are provided for inputs of the machine learning model, wherein the further training examples are labeled with target outputs with respect to a given task;

the further training examples are processed into outputs by the machine learning model);

a deviation of the outputs from the target outputs is assessed using a given cost function; and

parameters that characterize a behavior of the machine learning model are optimized with a goal that the assessment by the cost function is expected to improve during further processing of labeled training examples.

**25**. The method according to claim **22**, wherein the machine learning model is configured to ascertain a classification of the images or point clouds.

**26**. The method according to claim **24**, wherein the pre-trained parameters are retained during a training with the further training examples.

**27**. The method according to claim **24**, wherein the further training examples belong to a different distribution or different domain than the training examples used for the pre-training.

**28**. A non-transitory machine-readable data carrieron which is stored a computer program for unsupervised pre-training of a machine learning model, the computer program, when executed by more or more computers, causing the one or more computers to perform the following steps:

providing a set of training examples for inputs of the machine learning model;

specifying a region of the machine learning model that is to be pre-trained;

generating variations from each of the training examples;

processing each of the variations into a first output in a first processing branch, the first processing branch including at least one first instance of the region to be pre-trained;

processing each of the variations into a second output in a second processing branch, the second processing branch including at least one second instance of the region to be pre-trained;

for each of the variations, ascertaining a similarity of the first output generated from the variation to an aggregation of the second outputs generated from all of the other variations of the same training example;

optimizing parameters that characterize a behavior of the first instance of the region to be pre-trained, with a goal of maximizing the ascertained similarity;

specifying the optimized parameters as pre-trained parameters of the machine learning model.

**29**. One or more computers equipped by a non-transitory machine-readable data carrieron which is stored a computer program for unsupervised pre-training of a machine learning model, the computer program, when executed by the more or more computers, causing the one or more computers to perform the following steps:

providing a set of training examples for inputs of the machine learning model;

specifying a region of the machine learning model that is to be pre-trained;

generating variations from each of the training examples;

processing each of the variations into a first output in a first processing branch, the first processing branch including at least one first instance of the region to be pre-trained;

processing each of the variations into a second output in a second processing branch, the second processing branch including at least one second instance of the region to be pre-trained;

for each of the variations, ascertaining a similarity of the first output generated from the variation to an aggregation of the second outputs generated from all of the other variations of the same training example;

optimizing parameters that characterize a behavior of the first instance of the region to be pre-trained, with a goal of maximizing the ascertained similarity;

specifying the optimized parameters as pre-trained parameters of the machine learning model.

* * * * *