Internal distribution code:

(A) [-] Publication in OJ
(B) [-] To Chairmen and Members
(C) [-] To Chairmen
(D) [X] No distribution

Datasheet for the decision of 3 December 2024

Case Number: T 0080/23 - 3.5.06

Application Number: 15908513.3

Publication Number: 3370159

IPC: G06N99/00, G06F8/10, G06F15/16, G06F17/50

Language of the proceedings: EN

Title of invention:
MODEL PARAMETER FUSION METHOD AND APPARATUS

Applicant:
Huawei Technologies Co., Ltd.

Headword:
Model parameter combination/HUAWEI

Relevant legal provisions:
EPC Art. 56

Keyword:
Inventive step - (no)

Europäisches Patentamt
European Patent Office
Office européen des brevets

Beschwerdekammern

Boards of Appeal

Chambres de recours

Boards of Appeal of the
European Patent Office
Richard-Reitzner-Allee 8
85540 Haar
GERMANY
Tel. +49 (0)89 2399-0

Case Number: T 0080/23 - 3.5.06

D E C I S I O N
of Technical Board of Appeal 3.5.06
of 3 December 2024

Appellant: (Applicant)	Huawei Technologies Co., Ltd. Huawei Administration Building Bantian Longgang District Shenzhen, Guangdong 518129 (CN)
Representative:	Gill Jennings & Every LLP The Broadgate Tower 20 Primrose Street London EC2A 2ES (GB)
Decision under appeal:	**Decision of the Examining Division of the European Patent Office posted on 27 July 2022 refusing European patent application No. 15908513.3 pursuant to Article 97(2) EPC.**

Composition of the Board:

Chairman	M. Müller
Members:	M. Domingo Vecchioni
	B. Müller

Summary of Facts and Submissions

I. The appeal is against the decision of the examining division to refuse European patent application No. 15908513.3 because claim 1 of the sole request on file lacked an inventive step, Article 56 EPC. The examining division referred to:

D1: US 2013/0290223 A1.

II. With the statement of grounds of appeal, the appellant requested that the decision of the examining division be set aside and a patent be granted on the basis of the set of claims underlying the contested decision (which was re-filed with the statement of grounds of appeal).

III. In a communication pursuant to Article 15(1) RPBA, the board presented its preliminary opinion. Claim 1 appeared to allow an interpretation according to which the "nodes" referred to in the claim were of an abstract nature. Following this broad interpretation, claim 1 would lack an inventive step for encompassing a straightforward computer-implementation of a non-technical method. Even if claim 1 were interpreted more narrowly, it would appear to lack an inventive step for failing to solve a technical problem over a known technical infrastructure for distributed machine learning such as the one disclosed in D1. Furthermore, it was not apparent that the method disclosed in the description solved the technical problem mentioned therein.

IV. Oral proceedings were held on 3 December 2024. The appellant maintained the requests filed with the statement of grounds of appeal (see point II above). At the end of the oral proceedings, the chair announced the decision of the board.

V. Claim 1 reads as follows:

"A model parameter combination method, wherein the method includes: the model parameter training platform perform model parameter combination on data sets carrying data labels, to obtain a total model parameter, and the total model parameter is used to identify a data type of new data, wherein the model parameter training platform includes: the calculation server for iteration calculation; wherein the method is applied to a machine learning system, the machine learning system comprises at least one parameter collection group and at least one parameter delivery group, each parameter collection group is corresponding to at least one parameter delivery group, each parameter collection group comprises at least one node, each parameter delivery group comprises at least one node, a node comprised in the at least one parameter collection group is different from a node comprised in a corresponding parameter delivery group, and the method comprises:

when any parameter collection group meets an intra-group combination condition, combining model parameters of M nodes in the parameter collection group that meets the condition, to obtain a first model parameter of the parameter collection group that meets the condition, wherein a smallest quantity s of combination nodes in the parameter collection group that meets the condition ≤ M ≤ a total quantity of nodes comprised in the parameter collection group that meets the condition; and

sending, to N nodes in a parameter delivery group corresponding to the parameter collection group that meets the condition, the first model parameter of the parameter collection group that meets the condition, wherein $1 \leq N \leq$ a total quantity of nodes comprised in the parameter delivery group corresponding to the parameter collection group that meets the condition."

Reasons for the Decision

The application

1. The application relates to what is commonly referred to as distributed or federated machine learning.

A method is proposed for training a machine learning model (determining "model parameters") using a computer system ("model parameter training platform") comprising a plurality of "calculation servers" - also named "nodes" - which may be "common computers" (paragraphs [0114]-[0116] of the original description; fig. 1).

The training dataset ("original data for model parameter training") comprises labelled data (e.g. animal images carrying an animal label). Each node stores in local storage only a part of the whole training dataset and performs training based on that part (paragraphs [0115]-[0116]; fig. 2: steps 201 and 202).

The proposed method ("model parameter combination method") appears to specifically relate to how to "combine" the model parameters learned separately by each node so as to obtain - after a certain number of iterations of (local) training and combination - "final" or "total" model parameters. The final model parameters are provided for subsequent use of the model (paragraphs [0116]-[0118]).

2. Two prior art methods are acknowledged in the description: a "first method" involving a "parameter server" that collects and combines model parameters, and a "second method" involving a node sending its model parameters to another node where they are combined with the model parameters computed by the other node, this process being repeated from node to node (paragraph [0004]).

The description states that "the first method has a relatively high performance requirement for a parameter server configured to perform model parameter combination, and is prone to cause a shutdown, and the second method requires more data to be stored and a large data transmission volume" (paragraph [0004]).

3. Against this background, the proposed method is said to "resolve a problem that model parameter combination has a high performance requirement for a parameter server, and a large data transmission volume" (paragraph [0005]).

4. In the proposed method, the nodes are assigned to at least one "parameter collection group" (PCG) and at least one "parameter delivery group" (PDG). They are such that:
- each PCG and each PDG comprises at least one node,
- each PCG corresponds to at least one PDG,
- if a PCG corresponds to a PDG, there is at least one node in the former and one in the latter that are different from each other (paragraph [0118]).

An example is given in paragraphs [0202]-[0207] in

which 6 nodes numbered 0 to 5 are assigned to 2 PCGs and 3 PDGs as follows:
- PCG 0 = {0, 2, 4},
- PCG 1 = {1, 3, 5},
- PDG 0 = {0, 3},
- PDG 1 = {1, 4},
- PDG 2 = {2, 5}.

5. First, each node receives a part of the training dataset (fig. 1, step 201; paragraph [0119] "data subset from a data set") and trains the model ("performs iterative calculation") based on the respective "data subset" using an "initial model parameter" (step 202; paragraphs [0124]-[0126]).

When any parameter collection group meets an "intra-group condition" - defined as M nodes of the group having completed the current training, where M≥s for a preset minimum value s - the model parameters obtained by these M nodes are "combined" to obtain a "first model parameter" of the parameter collection group (step 203; paragraphs [0127]-[0128]).

It seems that whenever a single "model parameter" ("initial model parameter", "first model parameters") is mentioned it is actually referring to a plurality of parameters or to a vector of parameters (assuming the model involves a plurality of scalar parameters).

6. The description does not explain how (mathematically) the parameters are to be "combined".

It does however indicate that the combination may either be carried out by a "device independent of the parameter collection group" (a "parameter server") or by one or more of nodes of the parameter collection group (a "control node" performing the whole combination or several nodes performing it iteratively, the last node being in that case referred to as a "combination node"; paragraphs [0132]-[0141]).

7. The "first model parameter" obtained by this combination operation is sent to N nodes in a parameter delivery group corresponding to the parameter collection group (step 204; paragraph [0149]-[0150]). The sending may be carried out either in a broadcast or iterative manner (paragraph [0151]).

8. When W parameter collection groups meet an "inter-group combination condition" - which may be defined as a preset number of intra-group combinations having been performed in each of these groups, all model parameters of nodes in each of the W parameter collection groups are "separately combined" so as to obtain a "second model parameter" of each of these W parameter collection groups (step 205; paragraphs [0156]-[0159].

Then, the second model parameters of all parameter collection groups in the W parameter collection groups are "combined" to obtain a "third model parameter (step 206; paragraph [0165]).

It seems that these two combination steps differ from each other in that, in the former, a "separate combination" and, in the latter, an "overall combination" is performed (see e.g. paragraphs [0156], [0165] and [0178]). How these two types of combination actually differ from each other is unclear to the board.

In any case, the combinations in both steps may be performed either by an independent device or by a node. The performing entity is referred to as an "executor" (paragraphs [0160]-[0163], [0167]-[0172]).

9. The third model parameter is then sent to the nodes in the W parameter collection groups and to corresponding parameter delivery groups, in both cases either by broadcast or iteratively (paragraphs [0179]-[0180]).

10. When a "preset condition" is met, which may be that a specific time has passed, "nodes included in a parameter collection group and a parameter delivery group" are "regrouped" (step 207; paragraphs [0183]-[0184]).

After the "regrouping", steps 202 to 207 are repeated "based on a data subset and a current model parameter" until a "final model parameter" is output.

It is not entirely clear to the board what the "regrouping" means. From paragraphs [0186]-[0187], it seems to be an assigning of nodes to PCG and PDGs and of PCGs to PDGs. Perhaps it is a re-assignment different from those underlying the previous executions of steps 202 to 207.

11. The description asserts that thereby "a problem that model parameter combination has a high performance requirement for a parameter server, a large data transmission volume, and a dynamic calculation resource adjustment is resolved" (paragraph [0191]).

Inventive step

12. The examining division considered that the only technical features of claim 1 were a calculation server

and a machine learning system, known for instance from D1. The implemented method was by itself a non-technical, abstract mathematical method for "adapting parameters of an abstract model, using abstract parameter groups exchanged between abstract nodes of the model". According to T 49/99, information modelling was not a technical activity. It was not derivable from claim 1 that any technical effect is achieved by this method, in particular not "alleviating a requirement of high performance on a parameter server ... while reducing the amount of data transmission also", as had been argued by the appellant. This method did thus not solve a technical problem. The claimed implementation of this non-technical method on a system like that of D1 being obvious, claim 1 lacked an inventive step, Article 56 EPC (decision under appeal, points 1-3).

13. The appellant argued that the invention of claim 1 "does not relate to modelling per se, but rather [to] a method for structuring and operating a model parameter combination apparatus involving the conception and implementation of a complex system that unquestionably includes features of a technical character". The claimed method achieved the "advantageous technical effect of reducing the amount of processing and data transmission required with respect to the cited prior art". Such an effect was technical according to G 1/19, which had "superseded and replaced" T 49/99. Furthermore, the "intra-group [...] conditional combining of model parameters and the sending of the first model parameter to nodes in a parameter delivery group corresponding to the condition-meeting parameter collection group features" conferred technicality upon the claimed method. As D1 disclosed no teaching pointing towards the claimed solution, claim 1 was

inventive (statement of grounds of appeal, pages 1-4).

As to the examining division's argument that this effect was not derivable from claim 1, the appellant noted that "the skilled person in the fields of machine learning and model parameter combination would [...] be capable of straightforwardly working the invention based on the claim wording alone so as to solve the [...] formulated objective technical problem of how to provide an improved machine learning system and method, and would experience no doubt as to credibility of [sic] substantially all embodiments encompassed by that claim wording exhibit these performance-improving effects upon which the invention is based" (statement of grounds of appeal, page 3, first paragraph).

At the oral proceedings, the appellant argued that the invention was not about mathematics but concerned a practical computational problem, namely how to improve transmission efficiency and avoid server overload. This was solved by a conditional local combination of model parameters, independently of whether the "nodes" are realised as separate computers or as software on a single computer. The dynamic selection of nodes for the combination of parameters enabled to take into account that some nodes may become faulty and enabled also scalability.

14. The board considers that the subject-matter of claim 1 lacks an inventive step, Articles 52(1) and 56 EPC.

14.1 Claim 1 is far from reflecting the method proposed in the description, which has been outlined in points 1 to 11 above.

14.2 Claim 1 is directed to a "model parameter combination method" and refers, on the one hand, to a "machine learning system" and to "nodes", and, on the other hand, to a "model parameter training platform" that includes a (single) "calculation server". Claim 1 does not establish a clear relationship between these two pairs of entities.

The board considers that the terms "machine learning system" and the "nodes" used in claim 1 may be interpreted - in the context of that claim - as referring to an abstract system with abstract nodes "of the model", like a neural network or a graphical model with nodes.

While this is not how these terms are used in the description, where "nodes" refer to different "calculation servers" (see point 1 above), the board considers that claim 1 does not exclude such a broad interpretation. This is in particular the case as claim 1 refers to a plurality of nodes but only to a single "calculation server". Furthermore, at the oral proceedings, the appellant did not contest this broad interpretation of claim 1 and agreed that claim 1 encompassed a realisation of the whole method by a single computer ("the calculation server"), the "nodes" being all realised in software.

Hence, claim 1 encompasses embodiments in which a single technical entity - "the calculation server", which may be a conventional computer system - is used to carry out all the claimed method steps.

14.3 The remainder of claim 1 defines an abstract computation method, involving abstract "nodes" having "model parameters" (claim 1 does not even specify that the nodes perform computations) and exchanging parameters between them under certain conditions (again similar to how the operation of an abstract neural network or graphical model may be described).

The only operations mentioned in claim 1 as being part of the claimed method are "combining model parameters of M nodes in the parameter collection group" to "obtain a first model parameter" if "an intra-group condition" is met by any parameter collection group (a collection of nodes), and the "sending" of the "first model parameter" to "N nodes of in a parameter delivery group corresponding to the parameter collection group". The "intra-group condition" is defined by reference to a "condition" that is to be met by at least s individual nodes of a "parameter collection group", but that "condition" remains undefined in claim 1.

14.4 It is not apparent to the board that this abstract computation method contributes to solving a technical problem by producing a technical effect within the context of claim 1. In particular, the appellant's arguments in that respect (see point 13 above) are not convincing.

14.4.1 The computation of "model parameters" of an abstract "machine learning system", for instance by training based on an abstract data set, does not have any technical character. The complexity of the model is irrelevant in this regard, as is the question whether the resulting model is "improved".

14.4.2 Effects related to "high performance requirements" of a parameter server or "large transmission volume" cannot be derived from claim 1.

Claim 1 does not specify how the nodes are technically implemented and so no conclusions can be drawn from claim 1 on how the claimed operations affect "performance requirements" or "transmission volume". This is also the case because claim 1 does not specify under which "condition" a "first model parameter" is obtained by combination of model parameters of certain nodes and is then sent to other nodes, nor what the purpose of the nodes in the "parameter delivery group" is.

As these alleged effects are not derivable from claim 1, it can be left open whether they would qualify as technical effects following G 1/19.

14.4.3 Similarly, effects relating to taking into account the possibility of "faulty" nodes and to the scalability of the method are not derivable from claim 1, for failing to specify a particular technical implementation and to define the "condition" for combination of model parameters and the purpose of the "parameter delivery group".

14.5 The board judges therefore that claim 1 encompasses a straightforward technical implementation of an abstract computation method on a conventional computer system, where the abstract computation method makes no technical contribution in the context of claim 1 and thus cannot support the presence of an inventive step within the meaning of Article 56 EPC. It follows that the subject-matter of claim 1 lacks an inventive step.

15. The board notes *obiter* that even if the method were interpreted more narrowly in the light of the description, it would not appear to be derivable from it that the technical effects alleged by the appellant are achieved over a technical infrastructure for

distributed machine learning like the one disclosed in D1, for the reasons given in points 10.2 and 10.3 of the communication pursuant to Article 15(1) RPBA dated 14 November 2024.

Order

For these reasons it is decided that:

The appeal is dismissed.

The Registrar: The Chairman:

L. Stridde Martin Müller

Decision electronically authenticated