Internal distribution code:

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

Datasheet for the decision of 7 February 2024

Case Number: T 1425/21 - 3.5.06

Application Number: 19173347.6

Publication Number: 3557491

IPC: G06N3/08, G06N3/04, G06N20/20

Language of the proceedings: EN

Title of invention:
TRAINING DISTILLED MACHINE LEARNING MODELS

Applicant:
Google LLC

Headword:
Distilled Machine Learning Models/Google

Relevant legal provisions:
EPC Art. 56, 83

Keyword:
Inventive step - (no) - effect not made credible within the whole scope of claim

Decisions cited:
T 0702/20, T 0748/19

Catchword:

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
Fax +49 (0)89 2399-4465

Case Number: T 1425/21 - 3.5.06

D E C I S I O N
of Technical Board of Appeal 3.5.06
of 7 February 2024

Appellant: (Applicant)	Google LLC 1600 Amphitheatre Parkway Mountain View, CA 94043 (US)
Representative:	Marks & Clerk GST 1 New York Street Manchester M1 4HD (GB)
Decision under appeal:	**Decision of the Examining Division of the European Patent Office posted on 16 April 2021 refusing European patent application No. 19173347.6 pursuant to Article 97(2) EPC.**

Composition of the Board:

Chairman	M. Müller
Members:	T. Alecu
	A. Jimenez

Summary of Facts and Submissions

I. The appeal is against the decision of the Examining Division to refuse the application.

II. The Appellant requests that the decision be set aside and that a patent be granted on the basis of the main request or of one of six auxiliary requests. The main and the first five auxiliary requests correspond to those underlying the decision under appeal. The sixth auxiliary request was filed with the statement of grounds of appeal.

III. The main request and the second auxiliary request were refused for a lack of inventive step in view of

D1: X. ZENG, "Using a Neural Network to Approximate an Ensemble of Classifiers", NEURAL PROCESSING LETTERS, vol. 12, no. 3, 2000.

The other auxiliary requests were not admitted (see decision points 19 to 22) due to lack of compliance with Article 84 EPC or 123(2) EPC.

IV. In a communication accompanying a summons to oral proceedings, the Board informed the Appellant of its provisional opinion that claim 1 of all requests lacked inventive step in comparison with D1 or with any known machine learning method, because a technical effect could not be established.

V. During the oral proceedings, all requests were maintained. At the end of the oral proceedings, the chairman announced the Board's decision.

VI. Claim 1 of the main request defines:

A method performed by one or more computers, the method comprising:

training a first machine learning model, wherein the first machine learning model is configured to receive an input and generate a respective score for each of a plurality of classes representing a probability that the class is a classification of the input; and

training a second machine learning model on a plurality of training inputs, wherein the second machine learning model is also configured to receive inputs and generate scores for the plurality of classes representing a probability that the class is a classification of the input, wherein the second machine learning model has a different architecture to the first machine learning model and has fewer parameters than the first machine learning model such that generating output from the second machine learning model requires less memory than generating output from the first machine learning model, the training comprising:

processing each training input using the first machine learning model to generate a first target soft output for the training input; and

training the second machine learning model to, for each of the training inputs, generate a soft output that matches the first target soft output for the training input, wherein the first target soft output and the soft output that matches the first target soft output comprises a respective soft score for each of the plurality of classes generated by a last layer of the respective machine learning model, and wherein each soft score satisfies:

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

wherein q_i is the score for a class i, z_i is a weighted combination of outputs of a previous layer of the respective machine learning model received by the last layer, j ranges from 1 to a total number of classes in the plurality of classes, T is a temperature constant, and wherein to generate the soft scores, T is set to a higher value than used to generate scores for the classes after the respective machine learning model has been trained.

VII. Claim 1 of the first auxiliary request specifies in addition to that of the main request that

the storage requirements of the second machine learning model are lower than the storage requirements of the first machine learning model

and the step of

storing the second machine language model on a storage device associated with the one or more computers for subsequent deployment.

VIII. Claim 1 of the second auxiliary request specifies in addition to that of the main request the step of

deploying the second machine learning model on a device.

IX. Claim 1 of the third auxiliary request in comparison with that of the second auxiliary request qualifies the deploying step further by stating

wherein the device has too few memory resources for the deployment of the first machine learning model.

X. Claim 1 of the fourth auxiliary request specifies in addition to that of the second auxiliary request that

generating an output using the second machine learning model deployed on the device, and wherein generating the output using the second machine learning model requires less memory than would be required using the first machine learning model.

XI. Claim 1 of the fifth auxiliary request specifies in addition to that of the fourth auxiliary request that

the device is an embedded system.

XII. Claim 1 of the sixth auxiliary request corresponds to that of the fourth auxiliary request except that the deploying refers, instead to *"a device"* to

a computer of the one or more computers

and the generating step refers to *"the computer"* instead of *"the device".*

Reasons for the Decision

The application and the prior art

1. The application relates to machine learning models such as deep neural networks. It proposes to approximate "*cumbersome*" machine learning models with "*distilled*" machine learning models which require less computation

and/or memory when deployed. For instance the distilled model may be a neural network with fewer layers or fewer parameters. The cumbersome model may be an ensemble classifier, possibly combining full classifiers with specialist classifiers. The distilled model is trained on a "plurality of training inputs" and the associated outputs of the cumbersome model, so as to "*generate outputs that are not significantly less accurate than outputs generated by the cumbersome machine learning model*" (page 1).

1.1 The training procedure aims at minimizing the differences between the "soft outputs" of the distilled model and those of the cumbersome model on the given training inputs, e.g. through backpropagation (page 6).

1.2 The soft outputs represent a class probability obtained according to a form of the softmax equation using a "temperature" parameter T, which is set higher during training than during subsequent use (see description page 4, see claim 1 of all requests).

2. Document D1, like the application, proposes to replace an ensemble of classifiers with an approximator. In D1, the "*basic strategy is to approximate a given ensemble of classifiers by an alternative representation that needs much less storage, while still maintaining the same or similar accuracy as the ensemble*" (page 226, bottom).

2.1 The approximator in D1 is trained on what is called a "pseudo training set" labelled by the ensemble classifier and with the objective to minimize the difference between a class probability vector output by

the ensemble and the class probability vector output by the approximator (section 3.2).

2.2 D1 presents results (Table 1 on page 233) based on an ensemble of ten component classifiers where each component is a single-hidden-layer neural network, and so is the approximator.

Main request

3. The Examining Division acknowledged two differences of the claimed matter over D1 (decision, point 14.2 of the reasons). As paraphrased by the Appellant (statement of grounds of appeal point 2), these are that:

(a) the second machine learning model (distilled machine learning model) has fewer parameters than the first machine learning model (cumbersome machine learning model) such that generating output from the second machine learning model requires less memory than generating output from the first machine learning model; and
(b) the second machine learning model is trained based upon a soft score satisfying a particular form as set out in claim 1.

4. The Examining Division was of the opinion that these two differences did not provide a technical effect.

4.1 The training procedure based on a particular form of softmax was "*not based on technical considerations as regards the functioning of the one or more computers*" and did not serve a technical purpose (14.3).

4.2 Although it accepted that the distilled model required less storage, it argued that a technical effect could not be acknowledged on that basis because

(a) the claim allowed for both learning machines to co-exist on the same computer (14.4 and 14.5),
(b) no technical details of the device, on which the second machine learning model may run, were specified (14.4, but also 18.2).

The Appellant's arguments

5. The Appellant disagreed (statement of grounds of appeal, point 16): "*what is relevant is that the distinguishing features of claim 1 are motivated by technical considerations in that a second machine learning model is produced based on the first machine learning model that has a reduced memory requirement but is as effective as the first machine learning model through the use of feature (b)*". The effect of reducing memory usage was achieved independently of the device on which it was deployed.

6. In response to the Board's preliminary opinion that a technical effect could not be established, the Appellant made the following statements during the oral proceedings.

7. The goal of the claimed invention was to provide classification results using fewer resources: the smaller "distilled" model should classify as well as the larger "cumbersome" model. The claim itself did not cover any pair of learning models but it did define structural and functional limitations on them.

7.1 The learning models were limited structurally, as their outputs were obtained in a specific manner, namely as "soft outputs". These did not single out one class but represented a set of probability values for the different classes.

7.2 The learning models were also limited functionally,as the smaller model was claimed to output soft scores that "matched" those of the smaller one.

8. The process of training the smaller model on the basis of the outputs of the larger one transferred knowledge from the larger to the smaller model. That knowledge was the classification capability of the larger model, not its specific way of fitting the data. Therefore, the second, smaller, model, needed not be as complex as the first one, as it only needed to be able to produce equivalent output scores.

8.1 This transfer was enabled by the way the temperature parameter of the soft score outputs was set. It was larger during training than in use, so as to allow for learning, but to produce sharp results when used. This concept worked in principle for any types of models.

9. The Appellant also stated that the invention was made in an extremely fast moving field in which specific details of the models and their parameters would quickly be out of date. Accordingly, indicating such details and parameters was not useful and should not be required in a patent application.

9.1 The skilled person in this field was highly skilled, had an extensive knowledge of available architectures and types of networks, and would be able to select the most suitable ones for the task at hand. The skilled

person would also know, in view of the task at hand, which precision was required, i.e. when the outputs could be said to "match".

9.2 The Appellant acknowledged that the skilled person had to carry out some, but not an unreasonable number of experiments to select appropriate models. According to T 312/88, reasons 3.3, a small number of routine experiments was not an undue burden on the skilled person and not detrimental for patentability (albeit, in that case, with respect to sufficiency of disclosure).

10. The Appellant also argued that pairs of models which were unsuitable for the task were not within the scope of the claim, because the claims were limited to models with matching outputs. Accordingly, all models which *were* covered by the claim provided the technical effect.

11. The Appellant stated that, according to established case law of the boards of appeal, for a method of the type claimed, i.e. a computer implemented mathematical method, a technical effect could be acknowledged either because the method was applied to solve a specific technical problem, or because its implementation had to be considered technical. The latter was the case here. The claimed method provided for reduced memory use with "matching" (i.e. the same or equivalent) classification results, so that a technical effect was present.

12. The invention was defined by the specific form of the output - soft scores with temperature parameter - and the way the temperature parameter was used to obtain a reduced model with equivalent classification abilities. It was therefore justified to seek protection without a

limitation to any specific learning models or application contexts.

Claim interpretation

13. The Appellant interprets the feature of

"*training the second machine learning model to, for each of the training inputs, generate a soft output that matches the first target soft output for the training input*"

as a functional feature requiring *a selection* of a second model and its training *so that* its outputs *actually* "*match*", i.e. are the same or equivalent, to those of the first model.

14. The Board does not consider that the person skilled in the art would interpret the claims in that manner. Rather, in the Board's view, the skilled person would adopt the more immediate interpretation according to which the cited feature only defines the *objective* of the training process.

14.1 Indeed, a neural network is trained by modifying the values of its parameters so as to minimise an error functional which measures the difference between the desired and the actual output. The standard training proceeds sample by sample, and the parameters are optimised as a function of the error for each sample. This corresponds to the structure of the claim:

"..*the training comprising:*
processing each training input using the first machine learning model to generate a first target soft output for the training input; and

training the second machine learning model to, for each of the training inputs, generate a soft output that matches the first target soft output for the training input..".

15. In the Board's view, therefore, the claim does not require that the second model is selected so that its outputs, once trained, actually "match" those of the first one, but only that the training procedure has this "match" as an objective. There is no guarantee that this objective is reached.

Articles 52 and 56 EPC

16. The Board notes that the features differentiating the invention from D1, or even the entire set of features defining the distilled model and its training, as a difference to a known cumbersome learning model, are mathematical methods which cannot, under the established case law of the boards of appeal (the "COMVIK" approach), be taken into account for inventive step unless they contribute in a causal manner to a technical effect.

17. It accepts that the distilled model has reduced memory requirements when compared to the cumbersome model; after all this is expressly claimed. However, a reduction in storage or computational requirements of a machine learning model is insufficient, by itself, to establish a technical effect. One also has to consider the performance of the "reduced" learning model (see decision T 702/20, reasons 14.1, from this same Board).

18. It is not credible in general that any model with fewer parameters can be as accurate as the more complex one it is meant to replace. For example, the complexity or

architecture of the reduced model may be insufficient or inadequate for the given problem.

19. The Board disagrees with the Appellant's counter-argument that the invention (by "knowledge transfer" see point 8 above) reliably ensures that any given smaller network can provide the same accuracy as the given larger one. The input and output complexity is the same for both networks. Hence, also the smaller network must be complex enough to be able to model the input-output relationship (see e.g. D1, section 4.3, for a discussion on accuracy and complexity of approximating classifiers of a single type).

19.1 The Board also does not see that the temperature-based training process ensures that the smaller model has an equivalent accuracy. It is not clear how exactly the temperature must be first set (for both models), and then varied, and what accuracy may be expected. The application simply does not discuss this.

19.2 Since, in the Board's view, the claim does not imply a step of selecting or obtaining a smaller model, but simply defines one as a given, the Appellant's arguments relating to trial and error are not pertinent (and even if they were, they would not succeed, see below).

20. The Board concludes therefore that the technical effect advanced by the Appellant (see point 11 above) cannot be acknowledged over the whole scope of the claim, i.e. for all sets of smaller and larger models. The second model may use fewer resources, but it cannot be said to produce the same results and many smaller models will, in fact, be considerably worse.

20.1 In principle, it appears possible to argue that the smaller model represents a "good" trade-off between resource requirements and accuracy, i.e that the smaller model may be less accurate but have (predictably) smaller resource requirements. However, the application lacks any information in that regard.

20.2 Since no technical effect can be acknowledged, claim 1 of the main request lacks an inventive step.

Further remarks: the Appellant's claim interpretation and Article 83 EPC

21. As discussed with the Appellant during the oral proceedings, the Appellant's interpretation of the claim would give rise to an objection under Article 83 EPC. This objection, on which this decision does not rely, is presented here for the sake of argument.

22. Were the Board to adopt the interpretation of the Appellant, and assume that the claim implies a step of selecting a suitable smaller model, the Board disagrees with the argument that the skilled person would be able to provide smaller networks with reduced memory needs and equivalent accuracy with only "few routine tests" for all classification tasks.

23. The application itself does not guide the skilled person towards an understanding as to which distilled models might replace which cumbersome models, and how accurate they might be in comparison. No examples of pairs of cumbersome and distilled learning models are provided, nor are any results showing the performance of the distilled models.

23.1 While the skilled person might be aware of the various architectures and types of networks available from common general knowledge, the number of these possibilities is quite large. For each of them, downsizing can be done in different ways, by reducing the number of layers, of neurons, of weights etc., and each of these in various ways.

23.2 The trial-and-error process would also have to keep an eye on the desired trade-off between size and accuracy as already discussed above, which is not a simple endeavour. The performance of machine learning models is not easily predictable and can vary considerably according to the task at hand (see e.g. again D1).

24. Further, the Board does not see that the temperature-based training process as claimed simplifies the trial-and-error process. As already stated, the application simply does not discuss how exactly the temperature must be set and varied and what accuracy gains, if any, may be expected.

25. The Board further notes that the person skilled in the art must be able to carry out the invention over the whole scope of the claim, i.e. in principle for any classification task and given larger model. The claims cannot be, a priori, be construed as excluding instances which (e.g. after trial and error) turn out not to work (see T 748/19 reasons 13 to 13.2). Thus the argument of the Appellant that non-working embodiments do not fall under the scope of the claim cannot succeed.

25.1 The Board is therefore of the opinion that the application does not sufficiently teach how to carry out in practice the claimed invention. It rather only sketches

out an idea whose implementation over the full scope of the claim requires a separate research program.

Auxiliary requests

26. As is evident from the statement of grounds of appeal, the amendments in these requests were meant to ensure that the claim language actually implies the technical effect of reduced storage requirements for the distilled model vis-à-vis the cumbersome model. In the above analysis, the Board has assumed that such reduction is present in claim 1 of the main request but has nonetheless come to the conclusion that a technical effect cannot be acknowledged. The amendments of the auxiliary requests therefore cannot change the Board's conclusion.

26.1 During the oral proceedings, the Appellant did not challenge the Board's view on this point. Accordingly, the Board concludes that these requests lack inventive step as well.

Order

For these reasons it is decided that:

The appeal is dismissed.

The Registrar: The Chairman:

L. Stridde Martin Müller

Decision electronically authenticated