

MODEL OPTIMIZATION METHOD AND DEVICE, INFORMATION PROCESSING METHOD AND DEVICE

[0001] The present application claims priority to Chinese Patent Application No. 202411191438.2, titled "MODEL OPTIMIZATION METHOD AND DEVICE, INFORMATION PROCESSING METHOD AND DEVICE", filed on August 28, 2024 with the National Intellectual Property Administration, PRC, which is incorporated herein by reference in its entirety.

FIELD

[0002] Embodiments of the present disclosure relate to the technical field of machine learning, and in particular to a model optimization method and device, and an information processing method and device.

BACKGROUND

[0003] With the rapid growth of computing capability and data volume, Large Language Models (LLMs) have been widely developed and applied. These models have demonstrated outstanding performance in fields such as natural language processing, computer vision, and recommendation systems. The core technologies thereof include Transformer architecture, attention mechanism, and large-scale pre-training. However, as the scale of model parameters continues to expand, the computing cost and energy consumption of the model are becoming increasingly prominent. To address these issues, a Mixture of Experts (MoE) model has been proposed. The MoE model achieves sparsity in computation by dynamically selecting and activating some experts. Specifically, the MoE model includes one router and multiple experts, with each input activating only a small number of experts, significantly reducing the computing cost. In the conventional technology, although the MoE model performs well in reducing the computing cost, there is a problem of uncertainty in the expert selection process. These uncertainties are mainly manifested as some tokens with approximate scores when selecting experts, leading to an increase in randomness in the selection and thus affecting model performance. Therefore, an effective solution is urgently needed to solve the above

problem.

SUMMARY

[0004] In view of this, a model optimization method is provided according to an embodiment of the present disclosure. An information processing method, a model optimization device, an information processing device, a computing device, a computer-readable storage medium and a computer program product are further provided according to one or more embodiments of the present disclosure to solve the technical deficiencies existing in the conventional art.

[0005] In a first aspect of the embodiments according to the present disclosure, a model optimization method is provided, which includes:

inputting multiple segmentation units into a mixture of experts model, and calculating a target entropy corresponding to each of the multiple segmentation units by a router in the mixture of experts model;

dividing, based on the target entropy, the multiple segmentation units into global segmentation units and local segmentation units;

broadcasting the global segmentation units to all sub expert models in the mixture of experts model for processing, and broadcasting the local segmentation units to sub expert models that match the local segmentation units in the mixture of experts model for processing;

and

tuning parameters of the mixture of experts model based on processing results until obtaining a target mixture of experts model that meets an optimization stopping condition.

[0006] Preferably, the calculating a target entropy corresponding to each of the multiple segmentation units includes:

calculating, by the router in the mixture of experts model, a selection score for the segmentation unit to select each sub expert model in the mixture of experts model; and

determining an initial entropy of the selection score corresponding to each sub expert model, and obtaining the target entropy corresponding to the segmentation unit by normalizing the initial entropy.

[0007] Preferably, the dividing, based on the target entropy, the multiple segmentation units into global segmentation units and local segmentation units includes:

comparing the target entropy corresponding to each of the multiple segmentation units with a preset entropy threshold, selecting a segmentation unit with the target entropy greater than the preset entropy threshold based on a comparison result as a global segmentation unit,
5 and selecting a segmentation unit with the target entropy less than or equal to the preset entropy threshold as a local segmentation unit; or

sorting the multiple segmentation units based on the target entropy corresponding to each of the multiple segmentation units, to obtain a segmentation unit sequence, selecting the
10 global segmentation units from the segmentation unit sequence according to preset proportion information, and using remaining segmentation units except for the global segmentation units as the local segmentation units.

[0008] Preferably, the broadcasting the local segmentation units to sub expert models that match the local segmentation units in the mixture of experts model for processing includes:

15 calculating a target selection score for the local segmentation unit to select each sub expert model in the mixture of experts model; and

determining a local sub expert model that matches the local segmentation unit in the mixture of experts model based on the target selection score, and broadcasting the local segmentation unit to the local sub expert model for processing.

20 **[0009]** Preferably, the tuning parameters of the mixture of experts model based on processing results until obtaining a target mixture of experts model that meets an optimization stopping condition includes:

tuning the parameters of the mixture of experts model based on the processing results, and detecting whether the tuned mixture of experts model meets the optimization stopping
25 condition;

if the tuned mixture of experts model does not meet the optimization stopping condition, selecting multiple candidate segmentation units as the multiple segmentation units, and performing the step of inputting multiple segmentation units into a mixture of experts model, and calculating a target entropy corresponding to each of the multiple segmentation units by a

router in the mixture of experts model; and

if the tuned mixture of experts model meets the optimization stopping condition, using the tuned mixture of experts model as the target mixture of experts model.

[0010] Preferably, before the inputting multiple segmentation units into a mixture of experts model, the method further includes:

obtaining original sample data associated with the mixture of experts model, preprocessing the original sample data, to obtain multiple initial segmentation units; and

performing embedding processing on the multiple initial segmentation units to obtain multiple segmentation units.

[0011] In a second aspect of the embodiments according to the present disclosure, an information processing method is provided, which includes:

acquiring problem information inputted by a client, constructing multiple problem segmentation units based on the problem information, and inputting the multiple problem segmentation units into a target mixture of experts model, where the target mixture of experts model is obtained by the above method;

selecting a target sub expert model matching each of the multiple problem segmentation units in the target mixture of experts model to process the multiple problem segmentation units; and

combining prediction results of target sub expert models matching the multiple question segmentation units to obtain answer information, and feeding back the answer information to the client.

[0012] In a third aspect of the embodiments according to the present disclosure, a model optimization device is provided, which includes: a calculation module, a dividing module, a processing module and a parameter tuning module. The calculation module is configured to input multiple segmentation units into a mixture of experts model, and calculate a target entropy corresponding to each of the multiple segmentation units by a router in the mixture of experts model. The dividing module is configured to divide, based on the target entropy, the multiple segmentation units into global segmentation units and local segmentation units. The processing module is configured to broadcast the global segmentation units to all sub expert models in the mixture of experts model for processing, and broadcast the local segmentation

units to sub expert models that match the local segmentation units in the mixture of experts model for processing. The parameter tuning module is configured to tune parameters of the mixture of experts model based on processing results until a target mixture of experts model that meets an optimization stopping condition is obtained.

5 **[0013]** In a fourth aspect of the embodiments according to the present disclosure, an information processing device is provided, which includes: an information acquisition module, a segmentation processing module and an information feedback module. The information acquisition module is configured to acquire problem information inputted by a client, construct multiple problem segmentation units based on the problem information, and input
10 the multiple problem segmentation units into a target mixture of experts model, where the target mixture of experts model is obtained by the above method. The segmentation processing module is configured to select a target sub expert model matching each of the multiple problem segmentation units in the target mixture of experts model to process the multiple problem segmentation units. The information feedback module is configured to
15 combine prediction results of target sub expert models matching the multiple question segmentation units to obtain answer information, and feeding back the answer information to the client.

[0014] In a fifth aspect of the embodiments according to the present disclosure, a computing device is provided, which includes a memory and a processor. The memory is
20 configured to store computer executable instructions, the processor is configured to execute the computer executable instructions, where the computer executable instructions, when executed by the processor, perform the steps of the model optimization method or the information processing method described above.

[0015] In a sixth aspect of the embodiments according to the present disclosure, a
25 computer-readable storage medium is provided, which includes computer executable instructions stored thereon, where the computer executable instructions, when executed by a processor, perform the steps of the model optimization method or the information processing method described above.

[0016] In a seventh aspect of the embodiments according to the present disclosure, a
30 computer program product is provided, which includes computer programs or instructions, where the computer programs or instructions, when executed by a processor, perform the

steps of the model optimization method or the information processing method described above.

[0017] With the model optimization method according to the embodiment, in order to reduce the uncertainty of selection by the router and improve robustness of the model, so as to ensure prediction accuracy, during the model optimization stage, the multiple segmentation units may be inputted into the mixture of experts model, and the target entropy corresponding to each of the segmentation units may be calculated by the router in the mixture of experts model. On this basis, the multiple segmentation units may be divided into global segmentation units and local segmentation units based on the target entropy. Then, the global segmentation unit may be broadcasted to all sub expert models in the mixture of experts model for processing, and the local segmentation unit may be broadcasted to the sub expert models that match the local segmentation unit in the mixture of experts model for processing. Therefore, during the model optimization phase, uncertain segmentation units may be broadcasted to all sub expert models for learning, and certain segmentation units are still used for learning of the matching sub expert models, so as to tune parameters of the mixture of experts model based on the processing results to obtain the target mixture of experts model that meets the optimization stopping condition. In the fine-tuning process of the mixture of experts model, all sub expert models are required to learn the relevant knowledge of uncertain segmentation units to ensure better robustness of the sub expert model during the application stage.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] Figure 1 is a schematic diagram of a model optimization method according to an embodiment of the present disclosure;

[0019] Figure 2 is a flowchart of a model optimization method according to an embodiment of the present disclosure;

[0020] Figure 3 is a flowchart of an information processing method according to an embodiment of the present disclosure;

[0021] Figure 4 is a flowchart of a processing procedure of an information processing method according to an embodiment of the present disclosure;

[0022] Figure 5 is a schematic diagram of a structure of a model optimization device according to an embodiment of the present disclosure;

[0023] Figure 6 is a schematic diagram of a structure of an information processing device according to an embodiment of the present disclosure; and

5 [0024] Figure 7 is a block diagram of a structure of a computing device according to an embodiment of the present disclosure.

DETAILED DESCRIPTION OF EMBODIMENTS

10 [0025] Many specific details are set forth in the following description to facilitate a full understanding of the present disclosure. However, the present disclosure may be implemented in many other ways different from those described herein, and those skilled in the art may make similar generalizations without violating the connotations of the present disclosure, and thus the present disclosure is not limited by the specific embodiments described below.

15 [0026] Terms used in one or more embodiments of the present disclosure are used only for the purpose of describing a particular embodiment and are not intended to limit one or more embodiments of the present disclosure. The singular forms of "a", "the", and "said" used in one or more embodiments of the present disclosure and the claims are also intended to encompass the plural form, unless the context clearly indicates otherwise. It should also be understood that the term "and/or" used in one or more embodiments of the present disclosure
20 refers to and encompasses any or all possible combinations of one or more of the listed items in association.

[0027] It should be understood that while the terms "first", "second", etc. may be used in one or more embodiments of the present disclosure to describe various information, such information should not be limited to these terms. These terms are used only to distinguish the
25 same type of information from one another. For example, without departing from the scope of the one or more embodiments of the present disclosure, "a first..." may also be referred to as "a second...", and similarly, "a second..." may be referred to as "a first...". Depending on the context, the phrase "if" as used herein may be interpreted as "at the time of" or "when" or "in response to a determination".

30 [0028] In addition, it should be noted that the user information (including but not limited to

user device information, user personal information, etc.) and data (including but not limited to data used for analysis, stored data, displayed data, etc.) involved in one or more embodiments of the present disclosure are authorized by the user or sufficiently authorized by the parties, and that the set, use, and processing of the relevant data are subject to relevant laws, regulations, and regulations of relevant countries and regions, and a corresponding operation portal is provided for the user to choose to authorize or reject.

[0029] First, the terminologies involved in one or more embodiments of the present disclosure are explained below.

[0030] The Mixture of Experts (MoE) model refers to an ensemble learning method that combines multiple specialized sub models (i.e., "experts") to form a holistic model, with each "expert" contributing in their area of expertise.

[0031] Large Language Models (LLMs) are important models in the current field of Natural Language Processing (NLP). These models are typically based on deep learning techniques, particularly Transformer architectures such as extensions or variants of BERT, GPT. LLMs can process and understand complex natural language texts, perform generative tasks (such as text generation and dialogue generation), and analytical tasks (such as sentiment analysis and information extraction).

[0032] In the present disclosure, a model optimization method is provided. One or more embodiments of the present disclosure further relate to an information processing method, a model optimization device, an information processing device, a computing device, a computer-readable storage medium, and a computer program product, which will be described in detail in the following embodiments.

[0033] In practice, a standard MoE model selects several experts with the highest scores for calculation based on scores of a router. However, in practice, it is common for scores of the router to be approximate to each other, especially when dealing with complex data. These uncertain tokens have approximate scores when selecting experts, which makes it difficult to determine the best experts and may result in selecting the wrong experts, affecting performance of the model. Therefore, an effective solution is needed urgently to address the aforementioned issues.

[0034] Referring to the schematic diagram shown in Figure 1, with a model optimization

method according to this embodiment, in order to reduce uncertainty of selection by the router and improve robustness of the model, so as to ensure prediction accuracy, during the model optimization stage, multiple segmentation units may be inputted into a mixture of experts model, and target entropy corresponding to each of the segmentation units may be calculated by the router in the mixture of experts model. On this basis, the multiple segmentation units may be divided into global segmentation units and local segmentation units based on the target entropy. Then, the global segmentation unit may be broadcasted to all sub expert models in the mixture of experts model for processing, and the local segmentation unit may be broadcasted to the sub expert models that match the local segmentation unit in the mixture of experts model for processing. Therefore, during the model optimization phase, uncertain segmentation units may be broadcasted to all sub expert models for learning, and certain segmentation units are still used to the learning of matching sub expert models, so as to tune parameters of the mixture of experts model based on the processing results to obtain a target mixture of experts model that meets an optimization stopping condition. In the fine-tuning process of the mixture of experts model, all sub expert models are required to learn the relevant knowledge of uncertain segmentation units to ensure better robustness of the sub expert model during the application stage.

[0035] Refer to Figure 2, Figure 2 shows a flowchart of a model optimization method according to an embodiment of the present disclosure, which includes the following steps S202 to S208.

[0036] In step S202, multiple segmentation units are inputted into the mixture of experts model, and a target entropy corresponding to each of the segmentation units is calculated by a router in the mixture of experts model.

[0037] The model optimization method provided in this embodiment is applied to a fine-tuning task of the mixture of experts model (MoE model). After optimizing the mixture of experts model, if scores calculated by the router are approximate to each other, even if there is a problem of failing to determine the best sub expert model, it can still be ensured that the selected sub expert model provides an accurate prediction result, thereby effectively improving the robustness of the model. The mixture of experts model may be applied to any field, such as fields of natural language processing (the mixture of experts model is used for text classification, text generation, translation, etc.), computer vision (the mixture of experts

model is used for image classification, object detection, image generation, etc.), recommendation systems (the mixture of experts model is used for personalized recommendation), which is not limited in this embodiment. The model optimization method is described by taking a mixture of experts model for sentiment analysis of the inputted text as an example in this embodiment. For descriptions in other scenarios, reference may be made to the same or corresponding descriptions in this embodiment, which will not be repeated in this embodiment.

[0038] Specifically, the multiple segmentation units refer to multiple tokens obtained by segmenting a sample during the fine-tuning stage of the mixture of experts model, where a token may be a character unit and/or a word unit in the sample. Correspondingly, the mixture of experts model specifically refers to a MoE model including a router and multiple sub expert models. The router is used to calculate a sub expert model matching each of the segmentation units, so that the segmentation unit may be submitted to the matching sub expert model for processing during the application phase. Correspondingly, the target entropy specifically refers to a result obtained by calculating the entropy of each of the segmentation units by the router. The target entropy is used to measure certainty of the segmentation unit in selecting a sub expert model. A high entropy corresponds to a worse certainty, that is, high entropy reflects that the possibility of a sub expert model selected by a segmentation unit optimally matches the segmentation unit in the mixture of experts model is uncertain, which may be optimal or not. Therefore, it is required to broadcast this type of segmentation units to all sub expert models with the model optimization method according to this embodiment for relevant knowledge learning, to eliminate the problem of decreased prediction accuracy caused by the uncertainty, ensuring that, even if an uncertain token is assigned with an sub expert model, the sub expert model can still provide an accurate prediction result for the uncertain token in the application stage.

[0039] Based on this, in order to reduce the uncertainty of selection of the router, and improve robustness of the model to ensure prediction accuracy, during the model optimization stage, multiple segmentation units may be inputted into the mixture of experts model, and the router in the mixture of experts model may be used to calculate the target entropy corresponding to each of the segmentation units. The target entropy may be used to measure the certainty of each segmentation unit in selecting a sub expert model, which facilitates the subsequent division of segmentation units based on types, so as to fine tune the mixture of

experts model to meet the usage requirements.

[0040] Furthermore, in order to ensure that the mixture of experts model meets the requirements after fine-tuning, the segmentation units inputted into the mixture of experts model for relevant knowledge learning may be obtained by processing original sample data.

5 In this embodiment, the specific implementation is described as follows:

obtaining original sample data associated with the mixture of experts model, preprocessing the original sample data, to obtain multiple initial segmentation units; and performing embedding processing on the initial segmentation units to obtain multiple segmentation units.

10 **[0041]** Specifically, the original sample data refers to a text sample used for fine-tuning the mixture of experts model. Correspondingly, preprocessing specifically refers to a process of performing segmenting, encoding and the like on the original sample data. Correspondingly, embedding processing specifically refers to a process of converting an initial segmentation unit into a vector expression.

15 **[0042]** Based on this, after obtaining the original sample data associated with the mixture of experts model, the original sample data may be preprocessed first to generate multiple initial segmentation units based on a preprocessing result. On this basis, in order to be processed by the mixture of experts model, an embedding processing may be performed on each of the multiple initial segmentation units to obtain multiple segmentation units, which may be used
20 to fine tune the mixture of experts model subsequently.

[0043] In summary, by preprocessing the original sample data, it can be ensured that the sample data is converted into multiple segmentation units that meet the input requirements of the mixture of experts model, thereby achieving fine-tuning of the model.

[0044] Furthermore, in order to accurately reflect the certainty of the segmentation unit in
25 selecting each sub expert model through the target entropy, the target entropy corresponding to the segmentation unit may be calculated in the following way. In this embodiment, the calculating the target entropy corresponding to each of the multiple segmentation units includes:

calculating, by the router in the mixture of experts model, a selection score for the
30 segmentation unit to select each sub expert model in the mixture of experts model; and

determining an initial entropy of the selection score corresponding to each sub expert model, and obtaining the target entropy corresponding to the segmentation unit by normalizing the initial entropies.

[0045] Specifically, the selection score refers to a score of a selection probability calculated by the router for the segmentation unit to select each sub expert model. A high score corresponds to a high probability of selecting the sub expert model. Correspondingly, the initial entropy specifically refers to entropy calculated based on the selection score, which is used to indicate the certainty of selecting each sub expert model for the segmentation unit. The selection score is normalized, the target entropy that reflects the certainty of selecting the sub expert model for the segmentation unit may be obtained.

[0046] Based on this, for each segmentation unit, the selection score for the segmentation unit to select each sub model in the mixture of experts model may be calculated by the router in the mixture of experts model. On this basis, the initial entropy corresponding to the selection scores of each sub expert model may be determined first, and the initial entropies may be normalized, to merge all the initial entropies, thereby obtaining the target entropy that reflects the selection of the segmentation unit on the sub expert model, to facilitate the subsequent division of multiple segmentation units.

[0047] In practice, after calculating the selection score corresponding to the segmentation unit, the router may calculate the initial entropy according to the following formula (1):

$$H(x) = -\sum_i g_i \log(g_i) \quad (1)$$

where x represents the segmentation unit, g_i represents a selection score for the segmentation unit to select the i -th sub expert model, and $H(x)$ represents the initial entropy corresponding to the selection score.

[0048] After obtaining the initial entropy, the initial entropies corresponding to all sub expert models may be normalized to obtain a normalized entropy, i.e. the target entropy, which may be calculated according to the following formula (2):

$$H_{\text{norm}}(x) = [H(x)] / [\log(N)] \quad (2)$$

where $H_{\text{norm}}(x)$ represents the target entropy, and N represents the total number of sub expert models.

[0049] For example, in order to enhance the prediction effect of the MoE model for text sentiment analysis, the MoE model is fine tuned. The MoE model includes three sub expert models, where a sub expert model a is adept at processing texts containing positive vocabulary and outputting prediction results of positive emotions, a sub expert model b is adept at processing texts containing negative vocabulary and outputting prediction results of negative emotions, and a sub expert model c is adept at processing neutral texts and outputting neutral prediction results. On this basis, N tokens may be obtained and inputted into the MoE model. Scores for each token to select three sub expert models are calculated by a router in the MoE model, and it is determined that the score for a token1 to select a sub expert model a is S_{a1} , the score for the token1 to select a sub expert model b is S_{b1} , and the score for the token1 to select a sub expert model c is S_{c1} ...the score for a token2 to select the sub expert model a is S_{a2} ... the score for a token N to select a sub expert model c is S_{cN} . Afterwards, according to the above formulas (1) and (2), the normalized entropy H_1 of the token1 relative to all sub expert models... and the normalized entropy H_N of the tokenN relative to all sub expert models may be calculated, so as to further distinguish different types of tokens based on the normalized entropy corresponding to each token, to fine tune the MoE model.

[0050] In summary, the entropy is calculated to reflect the certainty of each segmentation unit in selecting each sub expert model, so that uncertain tokens may be selected, which facilitates broadcasting these tokens to all sub expert models for relevant knowledge learning during the fine-tuning stage, making the model more robust.

[0051] In step S204, the multiple segmentation units are divided into global segmentation units and local segmentation units based on the target entropy.

[0052] Specifically, after obtaining the target entropy corresponding to each segmentation unit, considering that the target entropy is used to reflect the certainty of each segmentation unit in selecting sub expert models, division may be performed based on the target entropy corresponding to each segmentation unit, to determine a segmentation unit with high uncertainty as a global segmentation unit, and determine a segmentation unit with low uncertainty as a local segmentation unit, which facilitates fine tuning of the mixture of experts model subsequently.

[0053] The global segmentation unit specifically refers to a segmentation unit with stronger

uncertainty in the multiple segmentation units, that is, when this type of segmentation unit has a stronger uncertainty when selecting its matching sub expert model through the router. Correspondingly, the local segmentation unit specifically refers to a segmentation unit that can accurately select its matching sub expert model from multiple segmentation units, and this type of segmentation unit has weak uncertainty.

[0054] Furthermore, after obtaining the target entropy corresponding to each segmentation unit as described above, the dividing into local segmentation units and global segmentation units is actually the division based on the certainty and uncertainty of segmentation units based on the entropy. In a specific implementation, the division may be implemented in the following two ways:

(1) comparing the target entropy corresponding to each of the multiple segmentation units with a preset entropy threshold, selecting a segmentation unit with the target entropy greater than the preset entropy threshold as a global segmentation unit based on a comparison result, and selecting a segmentation unit with the target entropy less than or equal to the preset entropy threshold as a local segmentation unit;

(2) sorting the multiple segmentation units based on the target entropy corresponding to each of the multiple segmentation units, to obtain a segmentation unit sequence, selecting the global segmentation units from the segmentation unit sequence according to preset proportion information, and using remaining segmentation units except for the global segmentation units as the local segmentation units.

[0055] Specifically, the preset entropy threshold refers to a threshold for selecting uncertain segmentation units and certain segmentation units, which may be set according to actual needs and is not limited in this embodiment. Correspondingly, the segmentation unit sequence specifically refers to a sequence obtained by sorting the multiple segmentation units according to the size of the target entropy. The preset proportion information specifically refers to a proportion of the selected number of segmentation units, which may be a specific quantity or a percentage, and is not limited in this embodiment.

[0056] Based on this, after obtaining the target entropy that can reflect the certainty of the segmentation unit in selecting a sub expert model, on the one hand, the target entropy corresponding to each segmentation unit may be compared with a preset entropy threshold. Based on a comparison result, a segmentation unit with a target entropy greater than the preset

entropy threshold may be selected as a global segmentation unit, and a segmentation unit with a target entropy less than or equal to the preset entropy threshold may be selected as a local segmentation unit, so that the fine-tuning of the mixture of experts model may be performed subsequently based on the global segmentation units and the local segmentation units.

5 **[0057]** On the other hand, the multiple segmentation units may be sorted based on their corresponding target entropies to obtain a segmentation unit sequence. On this basis, the global segmentation units may be selected from the segmentation unit sequence according to the preset proportion information, and the selection process may be completed in the order of the segmentation units in the sequence. Finally, the remaining segmentation units except for
10 the global segmentation units may be used as local segmentation units.

[0058] In specific implementation, the multiple segmentation units may be sorted in an order of target entropies from high to low to form the segmentation unit sequence. Based on this, the segmentation units of the preset proportion information may be selected in order from high to low as the global segmentation units.

15 **[0059]** In practice, the division of the segmentation units according to the target entropy may be performed by selecting a processing suitable for the current scenario according to actual needs, which is not limited in this embodiment.

[0060] Following the above example, in a case of obtaining the normalized entropy corresponding to each of the N tokens, the N tokens may further be sorted from high to low
20 according to the size of the normalized entropy, after which the tokens with entropies in the top 5% of the tokens are selected as the uncertain tokens, including {token1, token2, token4...}, and after which the N tokens other than the uncertain tokens are selected as certain tokens including {token3, token5, token6...}. Subsequently, the fine-tuning of the MoE model may be performed based on the uncertain tokens and the certain tokens.

25 **[0061]** In summary, by dividing into global segmentation units and local segmentation units in different ways, it can be ensured that the model can select different strategies to optimize for different types of segmentation units during fine-tuning stage of the model, thereby enhancing the prediction capability of the model.

[0062] In step S206, the global segmentation unit is broadcasted to all sub expert models in
30 the mixture of experts model for processing, and the local segmentation unit is broadcasted to

sub expert models that match the local segmentation unit in the mixture of experts model for processing.

[0063] Specifically, after determining the global segmentation units and the local segmentation units as described above, in order to solve the problem of inaccurate prediction result caused by inappropriate selection of sub expert models by the router for the segmentation unit, in the fine-tuning stage of the model, the global segmentation unit may be broadcasted to all sub expert models in the mixture of experts model for relevant knowledge learning, so that all sub expert models in the mixture of experts model can learn the relevant knowledge of the global segmentation unit, thereby ensuring that even there is a segmentation unit that cannot accurately match the sub expert models, the selected sub expert model can still provide an accurate prediction result in the application stage. In addition, the local segmentation unit may be broadcasted to the sub expert models that match the local segmentation unit in the mixture of experts model for processing, to normally fine-tune the sub expert models that match segmentation unit.

[0064] The sub expert model that matches the local segmentation unit specifically refers to a sub expert model that can accurately determine the processing of the local segmentation unit after calculating the selection score of the sub expert model by the router.

[0065] Furthermore, the local segmentation unit may be broadcasted to matching sub expert models by calculating selection scores. In this embodiment, the specific implementation is described as follows:

calculating a target selection score for the local segmentation unit to select each sub expert model in the mixture of experts model; and determining a local sub expert model that matches the local segmentation unit in the mixture of experts model based on the target selection score, and broadcasting the local segmentation unit to the local sub expert model for processing.

[0066] Specifically, the target selection score refers to a score for a local segmentation unit to select each sub expert model calculated by the router in the mixture of experts model. A high score corresponds to a high probability of being selected. Correspondingly, a local sub expert model specifically refers to a sub expert model that matches the local segmentation unit and is used to process the local segmentation unit.

[0067] Based on this, the standard MoE method may be used to select sub expert models for the local segmentation unit. That is, the target selection score for the local segmentation unit to select each sub expert model in the mixture of experts model may be calculated first. Afterwards, the local sub expert models that match the local segmentation unit may be determined in the mixture of experts model based on the target selection score, and the local segmentation unit may be broadcasted to the local sub expert models for processing.

[0068] Following the previous example, after obtaining uncertain tokens including {token1, token2, token4...}, and certain tokens including {token3, token5, token6...}, the uncertain tokens may be broadcasted to all sub expert models a, b, and c in the MoE model for learning.

For the certain tokens, the standard MoE method is used to select the sub expert model that matches each token for learning. The selection may be completed according to the scores calculated above. Therefore, learning results of all sub expert models can be integrated to update the model parameters subsequently.

[0069] In summary, the global segmentation unit is broadcasted to all sub expert models and the local segmentation unit is broadcasted to the matching sub expert models for processing, so that all sub expert models can learn relevant knowledge of uncertainty, thereby enabling the mixture of experts model to have more accurate prediction capabilities.

[0070] In step S208, the parameters of the mixture of experts model are tuned based on the processing result until a target mixture of experts model that meets an optimization stopping condition is obtained.

[0071] Specifically, in the above process, all sub expert models are used to process the global segmentation unit, and sub expert models that match the local segmentation unit are used to process the local segmentation unit, based on this, parameters of the mixture of experts model may be tuned according to the processing result to obtain the target mixture of experts model that meets the optimization stopping condition.

[0072] Furthermore, when tuning the parameters of the mixture of experts model, the target mixture of experts model that meets the requirements may be determined by comparing with an optimization stopping condition. In this embodiment, the specific implementation is described as follows:

tuning the parameters of the mixture of experts model based on the processing results,

and detecting whether the tuned mixture of experts model meets the optimization stopping condition; if the tuned mixture of experts model does not meet the optimization stopping condition, selecting multiple candidate segmentation units as the multiple segmentation units, and performing the step of inputting multiple segmentation units into a mixture of experts
5 model, and calculating a target entropy corresponding to each of the multiple segmentation units by a router in the mixture of experts model; and if the tuned mixture of experts model meets the optimization stopping condition, using the tuned mixture of experts model as the target mixture of experts model.

[0073] Based on this, when tuning the parameters of the mixture of experts model, it may
10 be determined whether the mixture of experts model meets the optimization stopping condition at each tuning stage. If the mixture of experts model does not meet the optimization stopping condition, it indicates that the model still needs to be trained continually, so multiple candidate segmentation units may be selected as multiple segmentation units, and the step of inputting multiple segmentation units into the mixture of experts model and calculating the
15 target entropy corresponding to each segmentation unit through the router in the mixture of experts model may be performed. Until a certain parameter tuning stage when the tuned model meets the optimization stopping condition, the tuned mixture of experts model may be used as the target mixture of experts model for deployment in business scenarios.

[0074] Following the previous example, after processing uncertain tokens by all sub expert
20 models and processing certain tokens by matching sub expert models, the learning results of all sub expert models a, b, and c may be combined to be used as a basis to update the parameters of the MoE model. When the updated MoE model meets the business deployment requirements, fine-tuning may be stopped and the updated MoE mode is deployed in the business scenario. After receiving the sentiment text to be analyzed inputted by the client, the
25 sentiment text may be inputted into the fine-tuned MoE model for processing, and sentiment information corresponding to the sentiment text to be analyzed may be determined for downstream business use.

[0075] With the model optimization method according to this embodiment, in order to
30 reduce the uncertainty of selection by the router and improve robustness of the model, so as to ensure prediction accuracy, during the model optimization stage, the multiple segmentation units may be inputted into the mixture of experts model, and the target entropy corresponding

to each of the segmentation units may be calculated by the router in the mixture of experts model. On this basis, the multiple segmentation units may be divided into global segmentation units and local segmentation units based on the target entropy. Then, the global segmentation unit may be broadcasted to all sub expert models in the mixture of experts model for processing, and the local segmentation unit may be broadcasted to the sub expert models that match the local segmentation unit in the mixture of experts model for processing. Therefore, during the model optimization phase, uncertain segmentation units may be broadcasted to all sub expert models for learning, and certain segmentation units are still used for learning of matching sub expert models, so as to tune parameters of the mixture of experts model based on the processing results to obtain the target mixture of experts model that meets the optimization stopping condition. In the fine-tuning process of the mixture of experts model, all sub expert models are required to learn the relevant knowledge of uncertain segmentation units to ensure better robustness of the sub expert model during the application stage.

[0076] Refer to Figure 3, Figure 3 shows a flowchart of an information processing method according to an embodiment of the present disclosure, which includes the following steps S302 to S306.

[0077] In step 302, problem information inputted by a client is acquired, multiple problem segmentation units are constructed based on the problem information, and the multiple problem segmentation units are inputted into a target mixture of experts model. The target mixture of experts model is obtained through the above method.

[0078] In step S304, the target sub expert model that matches each problem segmentation unit is selected in the target mixture of experts model to process the multiple problem segmentation units.

[0079] In step S306, the prediction results of the target sub expert models matching the question segmentation units are combined to obtain answer information, and the answer information is fed back to the client.

[0080] In the information processing method according to this embodiment, the training of the target mixture of experts model may be referred to the same or corresponding descriptions in the above embodiments, and is not repeated in this embodiment.

[0081] The information processing method according to the present disclosure is further described with reference to Figure 4 by using an example of the application of information processing method according to the present disclosure to a sentiment analysis scenario. Figure 4 shows a flowchart of a processing procedure of an information processing method according to an embodiment of the present disclosure, which specifically includes the following steps S402 to S420.

[0082] In step S402, multiple segmentation units are inputted into a mixture of experts model, and a target entropy corresponding to each of the segmentation units is calculated by a router in the mixture of experts model.

[0083] The calculating a target entropy corresponding to each of the segmentation units includes: calculating, by the router in the mixture of experts model, a selection score for the segmentation unit to select each sub expert model in the mixture of experts model; and determining an initial entropy of the selection score corresponding to each sub expert model, and obtaining the target entropy corresponding to the segmentation unit by normalizing initial entropies.

[0084] In step S404, the target entropy corresponding to each segmentation unit is compared with a preset entropy threshold, and a segmentation unit with a target entropy greater than the preset entropy threshold is selected as a global segmentation unit based on a comparison result.

[0085] In step S406, a segmentation unit with a target entropy less than or equal to the preset entropy threshold is selected as a local segmentation unit based on the comparison result.

[0086] In step S408, the global segmentation unit is broadcasted to all sub expert models in the mixture of experts model for processing.

[0087] In step S410, a target selection score for the local segmentation unit to select each sub expert model in the mixture of experts model is calculated.

[0088] In step S412, a local sub expert model that matches the local segmentation unit is determined based on the target selection score in the mixture of experts model, and the local segmentation unit is broadcasted to the local sub expert model for processing.

[0089] In step S414, the parameters of the mixture of experts model is tuned based on the

processing result until a target mixture of experts model that meets the optimization stopping condition is obtained.

[0090] In step S416, upon receiving the question information inputted by the client, multiple question segmentation units are constructed based on the question information and inputted into the target mixture of experts model.

[0091] In step S418, a target sub expert model that matches each problem segmentation unit is selected in the target mixture of experts model to process the multiple problem segmentation units.

[0092] In step S420, prediction results of the target sub expert models match the question segmentation units are combined to obtain answer information, and the answer information is fed back to the client.

[0093] In summary, in order to reduce the uncertainty of selection by the router and improve robustness of the model, so as to ensure prediction accuracy, during the model optimization stage, the multiple segmentation units may be inputted into the mixture of experts model, and the target entropy corresponding to each of the segmentation units may be calculated by the router in the mixture of experts model. On this basis, the multiple segmentation units may be divided into global segmentation units and local segmentation units based on the target entropy. Then, the global segmentation unit may be broadcasted to all sub expert models in the mixture of experts model for processing, and the local segmentation unit may be broadcasted to the sub expert models that match the local segmentation unit in the mixture of experts model for processing. Therefore, during the model optimization phase, uncertain segmentation units may be broadcasted to all sub expert models for learning, and certain segmentation units are still used for learning of the matching sub expert models, so as to tune parameters of the mixture of experts model based on the processing results to obtain the target mixture of experts model that meets the optimization stopping condition. In the fine-tuning process of the mixture of experts model, all sub expert models are required to learn the relevant knowledge of uncertain segmentation units to ensure better robustness of the sub expert model during the application stage.

[0094] Corresponding to the above method embodiments, an embodiment of a model optimization device is further provided according to the present disclosure. Figure 5 shows a schematic diagram of a structure of a model optimization device according to an embodiment

of the present disclosure. As shown in Figure 5, the device includes: a calculation module 502, a dividing module 504, a processing module 506 and a parameter tuning module 508. The calculation module 502 is configured to input multiple segmentation units into a mixture of experts model, and calculate a target entropy corresponding to each of the multiple segmentation units by a router in the mixture of experts model. The dividing module 504 is configured to divide, based on the target entropy, the multiple segmentation units into global segmentation units and local segmentation units. The processing module 506 is configured to broadcast the global segmentation units to all sub expert models in the mixture of experts model for processing, and broadcast the local segmentation units to sub expert models that match the local segmentation units in the mixture of experts model for processing. The parameter tuning module 508 is configured to tune parameters of the mixture of experts model based on processing results until a target mixture of experts model that meets an optimization stopping condition is obtained.

[0095] In an preferred embodiment, the target entropy corresponding to each of the multiple segmentation units is calculated by:

calculating, by the router in the mixture of experts model, a selection score for the segmentation unit to select each of the sub expert models in the mixture of experts model; and determining an initial entropy of the selection score corresponding to each of the sub expert models, and obtaining the target entropy corresponding to the segmentation unit by normalizing initial entropies.

[0096] In an preferred embodiment, the dividing module 504 is further configured to:

compare the target entropy corresponding to each of the multiple segmentation units with a preset entropy threshold, select a segmentation unit with the target entropy greater than the preset entropy threshold as a global segmentation unit based on a comparison result, and select a segmentation unit with the target entropy less than or equal to the preset entropy threshold as a local segmentation unit; or

sort the multiple segmentation units based on the target entropy corresponding to each of the multiple segmentation units, to obtain a segmentation unit sequence, select the global segmentation units from the segmentation unit sequence according to preset proportion information, and use remaining segmentation units except for the global segmentation units as the local segmentation units.

[0097] In an preferred embodiment, the processing module 506 is further configured to:

calculate a target selection score for the local segmentation unit to select each sub expert model in the mixture of experts model; and

determine a local sub expert model that matches the local segmentation unit in the mixture of experts model based on the target selection score, and broadcast the local segmentation unit to the local sub expert model for processing.

[0098] In an preferred embodiment, the tuning module 508 is further configured to:

tune the parameters of the mixture of experts model based on the processing results, and detect whether the tuned mixture of experts model meets the optimization stopping condition;

if the tuned mixture of experts model does not meet the optimization stopping condition, select multiple candidate segmentation units as the multiple segmentation units, and perform the step of inputting multiple segmentation units into a mixture of experts model, and calculating a target entropy corresponding to each of the multiple segmentation units by a router in the mixture of experts model; and

if the tuned mixture of experts model meets the optimization stopping condition, use the tuned mixture of experts model as the target mixture of experts model.

[0099] In an preferred embodiment, the device further includes:

a segmentation module, configured to original sample data associated with the mixture of experts model, preprocess the original sample data, to obtain multiple initial segmentation units; and perform embedding processing on the multiple initial segmentation units to obtain multiple segmentation units.

[0100] With the model optimization device according to this embodiment, in order to reduce the uncertainty of selection by the router and improve robustness of the model, so as to ensure prediction accuracy, during the model optimization stage, the multiple segmentation units may be inputted into the mixture of experts model, and the target entropy corresponding to each of the segmentation units may be calculated by the router in the mixture of experts model. On this basis, the multiple segmentation units may be divided into global segmentation units and local segmentation units based on the target entropy. Then, the global segmentation unit may be broadcasted to all sub expert models in the mixture of experts model for processing, and the local segmentation unit may be broadcasted to the sub expert

models that match the local segmentation unit in the mixture of experts model for processing. Therefore, during the model optimization phase, uncertain segmentation units may be broadcasted to all sub expert models for learning, and certain segmentation units are still used for learning of the matching sub expert models, so as to tune parameters of the mixture of experts model based on the processing results to obtain the target mixture of experts model that meets the optimization stopping condition. In the fine-tuning process of the mixture of experts model, all sub expert models are required to learn the relevant knowledge of uncertain segmentation units to ensure better robustness of the sub expert model during the application stage.

[0101] The above is a schematic solution of a model optimization device for this embodiment. It should be noted that the technical solution of the model optimization device belongs to the same concept as the technical solution of the model optimization method mentioned above. For details of the technical solution of the model optimization device that are not described in detail, please refer to the description of the technical solution of the model optimization method mentioned above.

[0102] Corresponding to the above method embodiments, an embodiment of an information processing device is further provided according to the present disclosure, and Figure 6 shows a schematic diagram of a structure of an information processing device according to an embodiment of the present disclosure. As shown in Figure 6, the device includes: an information acquisition module 602, a segmentation processing module 604 and an information feedback module 606. The information acquisition module 602 is configured to acquire problem information inputted by a client, construct multiple problem segmentation units based on the problem information, and input the multiple problem segmentation units into a target mixture of experts model, where the target mixture of experts model is obtained by the above method. The segmentation processing module 604 is configured to select a target sub expert model matching each of the multiple problem segmentation units in the target mixture of experts model to process the multiple problem segmentation units. The information feedback module 606 is configured to combine prediction results of target sub expert models matching the multiple question segmentation units to obtain answer information, and feeding back the answer information to the client.

[0103] A schematic solution of the information processing device of the embodiment is

described above. It should be noted that the technical solution of the information processing device and the technical solution of the information processing method described above belong to the same concept, and the details not described in detail in the technical solution of the information processing device may be referred to the description of the technical solution of the information processing method described above.

[0104] Figure 7 illustrates a block diagram of a structure of a computing device 700 according to an embodiment of the present disclosure. The components of the computing device 700 include, but are not limited to, a memory 710 and a processor 720. The processor 720 is connected to the memory 710 via a bus 730, and a database 750 is used to store data.

[0105] The computing device 700 further includes an access device 740, the access device 740 enables the computing device 700 to communicate via one or more networks 760. Examples of these networks include a public switched telephone network (PSTN), a local area network (LAN), a wide area network (WAN), a personal area network (PAN), or a combination of communication networks such as the Internet. The access device 740 may include one or more of any types of wired or wireless network interfaces (e.g., a network interface controller (NIC), such as an IEEE802.11 Wireless Local Area Network (WLAN) wireless interface, Worldwide Interoperability for Microwave Access (Wi-MAX) interface, Ethernet interface, Universal Serial Bus (USB) interface, cellular network interface, Bluetooth interface, Near Field Communication (NFC).

[0106] In an embodiment of the present disclosure, the above components of the computing device 700, as well as other components not shown in Figure 7, may also be connected to each other, such as via a bus. It should be understood that the block diagram of the structure of the computing device shown in Figure 7 is for exemplary purposes only and is not a limitation on the scope of the present disclosure. Those skilled in the art may add or replace other components as desired.

[0107] The computing device 700 may be any type of stationary or mobile computing device, including a mobile computer or a mobile computing device (e.g., a tablet computer, a personal digital assistant, a laptop computer, a notebook computer, a netbook, etc.), a cell phone (e.g., a smartphone), a wearable computing device (e.g., a smartwatch, smartglasses, etc.), or other types of mobile devices, or a stationary computing device such as desktop computer or a personal computer (PC). The computing device 700 may also be a mobile or

stationary server.

[0108] The processor 720 is used to execute computer executable instructions, which, when executed by the processor, perform the steps of the above-described model optimization method or information processing method.

5 **[0109]** A schematic solution of the computing device of the other embodiment is described above. It should be noted that the technical solution of the computing device and the technical solution of the model optimization method or the information processing method described above belong to the same concept, and the details not described in detail in the technical solution of the computing device may be referred to the description of the technical solution
10 of the model optimization method or the information processing method described above.

[0110] A computer-readable storage medium is further according to an embodiment of the present disclosure, which stores computer-executable instructions, the computer-executable instructions, when executed by a processor, perform the steps of the above-described model optimization method or information processing method.

15 **[0111]** A schematic solution of the computer-readable storage medium of the other embodiment is described above. It should be noted that the technical solution of the computer-readable storage medium and the technical solution of the model optimization method or the information processing method described above belong to the same concept, and the details not described in detail in the technical solution of the computer-readable
20 storage medium may be referred to the description of the technical solution of the model optimization method or the information processing method described above.

[0112] A computer program product is further according to an embodiment of the present disclosure, which includes computer programs or instructions, the computer programs or instructions, when executed by a processor, perform the steps of the above-described model
25 optimization method or information processing method.

[0113] A schematic solution of the computer program product of the other embodiment is described above. It should be noted that the technical solution of the computer program product and the technical solution of the model optimization method or the information processing method described above belong to the same concept, and the details not described
30 in detail in the technical solution of computer program product may be referred to the

description of the technical solution of the model optimization method or the information processing method described above.

[0114] Preferred embodiments of the present disclosure are described above. Other embodiments are within the scope of the claims. In some cases, the actions or steps described in the claims may be performed in a different order than in the embodiments and still achieve the desired results. In addition, the processes depicted in the drawings do not necessarily require the particular order or consecutive order shown to achieve the desired results. In some embodiments, multi-tasking and parallel processing is also possible or may be advantageous.

[0115] The computer instruction includes computer program codes, the computer program code may be in the form of source code, in the form of object code, in the form of an executable file or in some intermediate form or the like. The computer readable medium may include: any entity or device capable of carrying the computer program code, a recording medium, a USB disk, a removable hard disk, a diskette, an optical disk, a computer memory, a Read-Only Memory (ROM), a Random Access Memory (RAM), an electric carrier signal, a telecommunications signal, and software distribution medium, etc. It should be noted that the content of the computer-readable medium may be increased or decreased according to the requirements of the patent practice, for example, in some regions, according to the patent practice, the computer-readable medium does not include carrier signals and telecommunication signals.

[0116] It should be noted that, for each of the foregoing method embodiments, for the sake of simplicity of description, they are all expressed as a series of combinations of actions, but those skilled in the art should know that the embodiments of the present disclosure are not limited by the order of the described actions, because according to the embodiments of the present disclosure, certain steps may be performed in other orders or simultaneously. Secondly, those skilled in the art should also be aware that the embodiments described in the disclosure are preferred embodiments, and the actions and modules involved are not necessarily necessary for the embodiments of the present disclosure.

[0117] In the above-described embodiments, the description of each embodiment has its own focus, and the part not described in detail in a certain embodiment may be referred to the relevant description of other embodiments.

[0118] The above preferred embodiments of the present disclosure are intended only to aid

in the description of the present disclosure. The preferred embodiments are not an exhaustive recitation of all details, nor do they limit the invention to the specific embodiments described only. Apparently, many modifications and variations may be made in accordance with the contents of the embodiments of the present disclosure. These embodiments are selected and
5 specifically described in the present disclosure to better explain the principles and practice of the embodiments of the present disclosure so that those skilled in the art to which they belong will be able to well understand and utilize the present disclosure.

CLAIMS

1. A model optimization method, comprising:

inputting a plurality of segmentation units into a mixture of experts model, and
5 calculating a target entropy corresponding to each of the plurality of segmentation units by a router in the mixture of experts model;

dividing, based on the target entropy, the plurality of segmentation units into global segmentation units and local segmentation units;

broadcasting the global segmentation units to all sub expert models in the mixture of
10 experts model for processing, and broadcasting the local segmentation units to sub expert models that match the local segmentation units in the mixture of experts model for processing;
and

tuning parameters of the mixture of experts model based on processing results until obtaining a target mixture of experts model that meets an optimization stopping condition.

15 2. The model optimization method according to claim 1, wherein the calculating a target entropy corresponding to each of the plurality of segmentation units comprises:

calculating, by the router in the mixture of experts model, a selection score for the segmentation unit to select each of the sub expert models in the mixture of experts model; and

determining an initial entropy of the selection score corresponding to each of the sub
20 expert models, and obtaining the target entropy corresponding to the segmentation unit by normalizing initial entropies.

3. The model optimization method according to claim 1, wherein the dividing, based on the target entropy, the plurality of segmentation units into global segmentation units and local segmentation units comprises:

25 comparing the target entropy corresponding to each of the plurality of segmentation units with a preset entropy threshold, selecting a segmentation unit with the target entropy greater than the preset entropy threshold based on a comparison result as a global segmentation unit, and selecting a segmentation unit with the target entropy less than or equal to the preset entropy threshold as a local segmentation unit; or

sorting the plurality of segmentation units based on the target entropy corresponding to each of the plurality of segmentation units, to obtain a segmentation unit sequence, selecting the global segmentation units from the segmentation unit sequence according to preset proportion information, and using remaining segmentation units except for the global segmentation units as the local segmentation units.

4. The model optimization method according to claim 1, wherein the broadcasting the local segmentation units to sub expert models that match the local segmentation units in the mixture of experts model for processing comprises:

calculating a target selection score for the local segmentation unit to select each sub expert model in the mixture of experts model; and

determining a local sub expert model that matches the local segmentation unit in the mixture of experts model based on the target selection score, and broadcasting the local segmentation unit to the local sub expert model for processing.

5. The model optimization method according to claim 1, wherein the tuning parameters of the mixture of experts model based on processing results until obtaining a target mixture of experts model that meets an optimization stopping condition comprises:

tuning the parameters of the mixture of experts model based on the processing results, and detecting whether the tuned mixture of experts model meets the optimization stopping condition;

if the tuned mixture of experts model does not meet the optimization stopping condition, selecting a plurality of candidate segmentation units as the plurality of segmentation units, and performing the step of inputting a plurality of segmentation units into a mixture of experts model, and calculating a target entropy corresponding to each of the plurality of segmentation units by a router in the mixture of experts model; and

if the tuned mixture of experts model meets the optimization stopping condition, using the tuned mixture of experts model as the target mixture of experts model.

6. The model optimization method according to claim 1, wherein before the inputting a plurality of segmentation units into a mixture of experts model, the method further comprises:

obtaining original sample data associated with the mixture of experts model,

preprocessing the original sample data, to obtain a plurality of initial segmentation units; and
performing embedding processing on the plurality of initial segmentation units to obtain the plurality of segmentation units.

7. An information processing method, comprising:

5 acquiring problem information inputted by a client, constructing a plurality of problem segmentation units based on the problem information, and inputting the plurality of problem segmentation units into a target mixture of experts model, wherein the target mixture of experts model is obtained by the method according to any one of claims 1 to 7;

10 selecting a target sub expert model matching each of the plurality of problem segmentation units in the target mixture of experts model to process the plurality of problem segmentation units; and

combining prediction results of target sub expert models matching the plurality of question segmentation units to obtain answer information, and feeding back the answer information to the client.

15 8. A model optimization device, comprising:

a calculation module, configured to input a plurality of segmentation units into a mixture of experts model, and calculate a target entropy corresponding to each of the plurality of segmentation units by a router in the mixture of experts model;

20 a dividing module, configured to divide, based on the target entropy, the plurality of segmentation units into global segmentation units and local segmentation units;

a processing module, configured to broadcast the global segmentation units to all sub expert models in the mixture of experts model for processing, and broadcast the local segmentation units to sub expert models that match the local segmentation units in the mixture of experts model for processing; and

25 a parameter tuning module, configured to tune parameters of the mixture of experts model based on processing results until a target mixture of experts model that meets an optimization stopping condition is obtained.

9. An information processing device, comprising:

an information acquisition module, configured to acquire problem information inputted

by a client, construct a plurality of problem segmentation units based on the problem information, and input the plurality of problem segmentation units into a target mixture of experts model, wherein the target mixture of experts model is obtained by the method according to any one of claims 1 to 7;

5 a segmentation processing module, configured to select a target sub expert model matching each of the plurality of problem segmentation units in the target mixture of experts model to process the plurality of problem segmentation units; and

 an information feedback module, configured to combine prediction results of target sub expert models matching the plurality of question segmentation units to obtain answer
10 information, and feeding back the answer information to the client.

 10. A computing device, comprising:

 a memory, configured to store computer executable instructions; and

 a processor, configured to execute the computer executable instructions, wherein the computer executable instructions, when executed by the processor, perform the steps of the
15 method according to any one of claims 1 to 7.

 11. A computer-readable storage medium, comprising computer executable instructions stored thereon, wherein the computer executable instructions, when executed by a processor, perform the steps of the method according to any one of claims 1 to 7.

 12. A computer program product, comprising computer programs or instructions,
20 wherein the computer programs or instructions, when executed by a processor, perform the steps of the method according to any one of claims 1 to 7.