

模型优化方法及装置、信息处理方法及装置

本申请要求于 2024 年 8 月 28 日提交中国专利局、申请号为 202411191438.2、发明名称为“模型优化方法及装置、信息处理方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本说明书实施例涉及机器学习技术领域，特别涉及模型优化方法及装置、信息处理方法及装置。

背景技术

随着计算能力和数据量的快速增长，大型语言模型（LLMs）得到了广泛的发展和应用。这些模型在自然语言处理、计算机视觉、推荐系统等领域展现出了卓越的性能。其核心技术包括 Transformer 架构、注意力机制和大规模预训练。然而，随着模型参数规模的不断扩大，模型的计算成本和能耗问题也日益凸显。为了解决这一问题，提出了混合专家（Mixture of Experts, MoE）模型。MoE 模型通过动态选择和激活部分专家，从而实现了计算的稀疏性。具体而言，MoE 模型由一个路由器和多个专家组成，每个输入仅激活少量的专家，显著降低了计算成本。现有技术中，MoE 模型虽然在减少计算成本方面表现出色，但在专家选择过程中存在不确定性问题。这些不确定性主要表现为一些 tokens 在选择专家时得分接近，导致选择的随机性增加，从而影响模型性能。因此亟需一种有效的方案以解决上述问题。

发明内容

有鉴于此，本说明书实施例提供了一种模型优化方法。本说明书一个或者多个实施例同时涉及一种信息处理方法，一种模型优化装置，一种信息处理装置，一种计算设备，一种计算机可读存储介质以及一种计算机程序产品，以解决现有技术中存在的技术缺陷。

根据本说明书实施例的第一方面，提供了一种模型优化方法，包括：

将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值；

根据所述目标熵值将所述多个分词单元划分为全局分词单元和局部分词单元；

将所述全局分词单元广播至所述混合专家模型中的全部专家子模型进行处理，以及将所述局部分词单元广播至所述混合专家模型中匹配所述局部分词单元的专家子模型进行处理；

根据处理结果对所述混合专家模型进行调参，直至获得满足优化停止条件的目标混合专家模型。

可选地，所述多个分词单元中的任意一个分词单元对应的目标熵值的计算，包括：

通过所述混合专家模型中的路由器，计算分词单元选择所述混合专家模型中每个专家子模型的选择分值；

确定每个专家子模型对应的选择分值的初始熵值，并通过对所述初始熵值进行归一化处理，获得所述分词单元对应的目标熵值。

可选地，所述根据所述目标熵值将所述多个分词单元划分为全局分词单元和局部分词单元，包括：

将每个分词单元对应的目标熵值与预设熵值阈值进行比较，根据比较结果选择大于所述预设熵值阈值的分词单元作为全局分词单元，以及选择小于等于所述预设熵值阈值的分词单元作为局部分词单元；或者，

根据每个分词单元对应的目标熵值对所述多个分词单元进行排序，获得分词单元序列，按照预设占比信息在所述分词单元序列中选择全局分词单元，并将除所述全局分词单元外的剩余分词单元作为局部分词单元。

可选地，所述将所述局部分词单元广播至所述混合专家模型中匹配所述局部分词单元的专家子模型进行处理，包括：

计算所述局部分词单元选择所述混合专家模型中各个专家子模型的目标选择分值；

根据所述目标选择分值在所述混合专家模型中确定匹配所述局部分词单元的局部专家子模型，并将所述局部分词单元广播至所述局部专家子模型进行处理。

可选地，所述根据处理结果对所述混合专家模型进行调参，直至获得满足优化停止条件的目标混合专家模型，包括：

根据处理结果对所述混合专家模型进行调参，并检测调参后的混合专家模型是否满足优化停止条件；

若否，选择多个候选分词单元作为所述多个分词单元，并执行将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值步骤；

若是，将调参后的混合专家模型作为目标混合专家模型。

可选地，所述将多个分词单元输入至混合专家模型步骤执行之前，还包括：

获取关联所述混合专家模型的原始样本数据，对所述原始样本数据进行预处理，获得多个初始分词单元；

对所述多个初始分词单元分别进行嵌入处理，获得多个分词单元。

根据本说明书实施例的第二方面，提供了一种信息处理方法，包括：

获取客户端提交的问题信息，根据所述问题信息构建多个问题分词单元并输入至目标混合专家模型，其中，所述目标混合专家模型通过上述方法获得；

在所述目标混合专家模型中选择每个问题分词单元匹配的目标专家子模型对所述多个问题分词单元进行处理；

将每个问题分词单元匹配的目标专家子模型的预测结果进行合并，获得答案信息，并将所述答案信息反馈至所述客户端。

根据本说明书实施例的第三方面，提供了一种模型优化装置，包括：

计算模块，被配置为将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值；

划分模块，被配置为根据所述目标熵值将所述多个分词单元划分为全局分词单元和局部分词单元；

处理模块，被配置为将所述全局分词单元广播至所述混合专家模型中的全部专家子模型进行处理，以及将所述局部分词单元广播至所述混合专家模型中匹配所述局部分词单元的专家子模型进行处理；

调参模块，被配置为根据处理结果对所述混合专家模型进行调参，直至获得满足优化停止条件的目标混合专家模型。

根据本说明书实施例的第四方面，提供了一种信息处理装置，包括：

获取信息模块，被配置为获取客户端提交的问题信息，根据所述问题信息构建多个问题分词单元并输入至目标混合专家模型，其中，所述目标混合专家模型通过上述方法获得；

处理分词模块，被配置为在所述目标混合专家模型中选择每个问题分词单元匹配的目标专家子模型对所述多个问题分词单元进行处理；

反馈信息模块，被配置为将每个问题分词单元匹配的目标专家子模型的预测结果进行合并，获得答案信息，并将所述答案信息反馈至所述客户端。

根据本说明书实施例的第五方面，提供了一种计算设备，包括：
存储器和处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指令，该计算机可执行指令被处理器执行时实现上述模型优化方法或信息处理方法的步骤。

根据本说明书实施例的第六方面，提供了一种计算机可读存储介质，其存储有计算机可执行指令，该指令被处理器执行时实现上述模型优化方法或信息处理方法的步骤。

根据本说明书实施例的第七方面，提供了一种计算机程序产品，包括计算机程序或指令，该计算机程序或指令被处理器执行时实现上述模型优化方法或信息处理方法的步骤。

本实施例提供的模型优化方法，为了能够降低路由器选择不确定性，提高模型鲁棒性，从而保证预测准确性，可以在模型优化阶段，通过将多个分词单元输入至混合专家模型后，先利用混合专家模型中的路由器计算每个分词单元对应的目标熵值；在此基础上，可以根据目标熵值将多个分词单元划分为全局分词单元和局部分词单元；而后可以将全局分词单元广播至混合专家模型中的全部专家子模型进行处理，同时将局部分词单元广播至混合专家模型中匹配局部分词单元的专家子模型进行处理；实现在模型优化阶段，可以将不确定分词单元广播给全部专家子模型进行学习，而确定分词单元依旧用于匹配专家子模型的学习，从而实现根据处理结果对混合专家模型进行调参，以获得满足优化停止条件的目标混合专家模型。实现在混合专家模型的微调过程中，让全部专家子模型学习不确定分词单元的相关知识，以保证在应用阶段，专家子模型具有更好的鲁棒性。

附图说明

图1是本说明书一个实施例提供的一种模型优化方法的示意图；

图2是本说明书一个实施例提供的一种模型优化方法的流程图；

图3是本说明书一个实施例提供的一种信息处理方法的流程图；

图4是本说明书一个实施例提供的一种信息处理方法的处理过程流程图；

图5是本说明书一个实施例提供的一种模型优化装置的结构示意图；

图6是本说明书一个实施例提供的一种信息处理装置的结构示意图；

图7是本说明书一个实施例提供的一种计算设备的结构框图。

具体实施方式

在下面的描述中阐述了很多具体细节以便于充分理解本说明书。但是本说明书能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本说明书内涵的情况下做类似推广，因此本说明书不受下面公开的具体实施的限制。

在本说明书一个或多个实施例中所使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本说明书一个或多个实施例中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

应当理解，尽管在本说明书一个或多个实施例中可能采用术语第一、第二等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本说明书一个或多个实施例范围的情况下，第一也可以被称为第二，类似地，第二也可以被称为第一。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

此外，需要说明的是，本说明书一个或多个实施例所涉及的用户信息(包括但不限于用户设备信息、用户个人信息等)和数据(包括但不限于用于分析的数据、存储的数据、展示的数据等)，均为经用户授权或者经过各方充分授权的信息和数据，并且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准，并提供有相应的操作入口，供用户选择授权或者拒绝。

首先，对本说明书一个或多个实施例涉及的名词术语进行解释。

混合专家(MoE)模型：全称为 Mixture of Experts，是一种集成学习方法，通过将多个专业化的子模型(即“专家”)组合起来，形成一个整体模型，每个“专家”都在其擅长的领域内做出贡献。

LLMs：(Large Language Models)指的是大型语言模型，是当前自然语言处理(NLP)领域的一类重要模型。这些模型通常基于深度学习技术，特别是 Transformer 架构，如 BERT、GPT 等模型的扩展或变体。LLMs 能够处理和理解复杂的自然语言文本，进行生成式任务(如文本生成、对话生成)和分析式任务(如情感分析、信息抽取)。

在本说明书中，提供了一种模型优化方法。本说明书一个或者多个实施例同时涉及一种信息处理方法，一种模型优化装置，一种信息处理装置，一种计算设备，一种计算机可读存储介质以及一种计算机程序产品，在下面的实施例中逐一进行详细说明。

实际应用中，标准的 MoE 模型通过路由器的得分选择若干个得分最高的专家进行计算。然而，在实际应用中，路由器得分接近的情况较为常见，特别是在处理复杂数据时。这些不确定 tokens 在选择专家时的得分接近，导致无法确定最佳专家，从而可能选择错误的专家，影响模型性能。因此亟需一种有效的方案以解决上述问题。

参见图 1 所示的示意图，本实施例提供的模型优化方法，为了能够降低路由器选择不确定性，提高模型鲁棒性，从而保证预测准确性，可以在模型优化阶段，通过将多个分词单元输入至混合专家模型后，先利用混合专家模型中的路由器计算每个分词单元对应的目标熵值；在此基础上，可以根据目标熵值将多个分词单元划分为全局分词单元和局部分词单元；而后可以将全局分词单元广播至混合专家模型中的

全部专家子模型进行处理，同时将局部分词单元广播至混合专家模型中匹配局部分词单元的专家子模型进行处理；实现在模型优化阶段，可以将不确定分词单元广播给全部专家子模型进行学习，而确定分词单元依旧用于匹配专家子模型的学习，从而实现根据处理结果对混合专家模型进行调参，以获得满足优化停止条件的目标混合专家模型。实现在混合专家模型的微调过程中，让全部专家子模型学习不确定分词单元的相关知识，以保证在应用阶段，专家子模型具有更好的鲁棒性。

参见图 2，图 2 示出了根据本说明书一个实施例提供的一种模型优化方法的流程图，具体包括以下步骤。

步骤 S202，将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值。

本实施例提供的模型优化方法应用于对混合专家模型（MoE 模型）的微调任务，实现对混合专家模型进行优化后，如果遇到路由器计算分值接近的情况下，即使出现无法确定最佳专家子模型的问题，依旧可以保证被选择的专家子模型给出更为精准的预测结果，从而有效的提高模型鲁棒性。其中，混合专家模型可以是应用于任意领域的 MoE 模型，比如自然语言处理领域（混合专家模型用于文本分类、文本生成、翻译等），计算机视觉领域（混合专家模型用于图像分类、目标检测、图像生成等），推荐系统（混合专家模型用于个性化推荐）等，本实施例在此不做任何限定。本实施例以混合专家模型用于对输入文本进行情感分析为例对模型优化方法进行说明，其他场景中的描述均可参见本实施例中相同或相应的描述，本实施例在此不做过多赘述。

具体的，多个分词单元具体是指对混合专家模型进行微调阶段，对样本进行分词处理后得到的多个 token，其中，token 可以是样本中的字单元和/或词单元。相应的，混合专家模型具体是指包含路由器和多个专家子模型的 MoE 模型，其中，路由器用于计算每个分词单元所匹配的专家子模型，实现在应用阶段可以将分词单元提交至匹配的专家子模型进行处理。相应的，目标熵值具体是指利用路由器对每个分词单元进行熵值计算后得到的结果，目标熵值用于衡量分词单元选择专家子模型的确定性，熵值越高表明确定性越差，即：体现分词单元选择专家子模型后，该专家子模型是混合专家模型中匹配分词单元最佳的可能性不确定，可能是最佳，也可能不是，因此需要通过本实施例提供的模型优化方法，将该类分词单元广播至所有专家子模型进行相关知识学习，从而消除不确定性带来的预测精度下降问题，保证在应用阶段，如果遇到不确定的 token，为其分配专家子模型后，可以保证专家子模型针对不确定 token 依旧可以给出准确的预测结果。

基于此，为了能够降低路由器选择不确定性，提高模型鲁棒性，从而保证预测准确性，可以在模型优化阶段，通过将多个分词单元输入至混合专家模型后，先利用混合专家模型中的路由器计算每个分词单元对应的目标熵值；以通过目标熵值衡量每个分词单元选择专家子模型的确定性，进而便于后续进行分词单元类型划分，以将混合专家模型微调为满足使用需求的目标混合专家模型。

进一步的，为了能够保证混合专家模型微调后满足需求，输入混合专家模型用于学习相关知识的分词单元可以通过对原始样本数据进行处理后得到。本实施例中，具体实现方式如下：

获取关联所述混合专家模型的原始样本数据，对所述原始样本数据进行预处理，获得多个初始分词

单元；对所述多个初始分词单元分别进行嵌入处理，获得多个分词单元。

具体的，原始样本数据具体是指用于对混合专家模型进行微调的文本样本。相应的，预处理具体是指对原始样本数据进行分词、编码等处理的过程。相应的，嵌入处理具体是指对初始分词单元转换为向量表达的处理。

5 基于此，获取关联混合专家模型的原始样本数据后，可以先对原始样本数据进行预处理，以根据预处理结果生成多个初始分词单元；在此基础上，为了能够实现输入混合专家模型后进行处理，可以再对多个初始分词单元分别进行嵌入处理，从而获得多个分词单元，实现后续可以结合多个分词单元对混合专家模型进行微调。

10 综上，通过对原始样本数据进行预处理，可以保证样本数据转换为满足混合专家模型输入要求的多个分词单元，进而能够达到模型微调目的。

更进一步的，在计算分词单元对应的目标熵值时，为了能够通过目标熵值精准体现分词单元选择每个专家子模型的确定性，可以通过如下方式实现。本实施例中，所述多个分词单元中的任意一个分词单元对应的目标熵值的计算，包括：

15 通过所述混合专家模型中的路由器，计算分词单元选择所述混合专家模型中每个专家子模型的选择分值；确定每个专家子模型对应的选择分值的初始熵值，并通过对所述初始熵值进行归一化处理，获得所述分词单元对应的目标熵值。

20 具体的，选择分值具体是指利用路由器计算分词单元选择每个专家子模型的选择可能性分值，该分值越高，表明专家子模型被选择的概率越高。相应的，初始熵值具体是指根据选择分值计算得到的熵值，用于表明分词单元选择每个专家子模型的确定性，通过对其进行归一化处理，即可得到体现分词单元选择专家子模型的确定性的目标熵值。

基于此，针对任意一个分词单元，可以先通过混合专家模型中的路由器，计算分词单元选择所述混合专家模型中每个专家子模型的选择分值；在此基础上，可以先确定每个专家子模型对应的选择分值的初始熵值，再通过对初始熵值进行归一化处理，即可实现将全部初始熵值进行合并，从而获得体现分词单元选择专家子模型的目标熵值，以便于后续对多个分词单元进行划分。

25 实际应用中，路由器在计算得到分词单元对应的选择分值后，可以通过下述公式（1）计算初始熵值：

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

其中， x 表示分词单元， g_i 表示分词单元选择第 i 个专家子模型的选择分值， $H(x)$ 表示对应该选择分值的初始熵值。

30 在得到初始熵值后，可以将所有专家子模型对应的初始熵值进行归一化，进而得到归一化熵值，即目标熵值，其可以通过如下公式（2）计算获得：

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

其中， $H_{\text{norm}}(x)$ 表示目标熵值， N 表示专家子模型的总数量。

举例说明，为了使得对文本情感分析的 MoE 模型具备更强的预测效果，将对 MoE 模型进行微调；

其中，MoE模型包含3个专家子模型，专家子模型a擅长处理包含积极词汇的文本，输出积极情感的预测结果；专家子模型b擅长处理包含消极词汇的文本，输出消极情感的预测结果；专家子模型c擅长处理中性的文本，输出中性预测结果。在此基础上，可以获取N个token，并将N个token输入至MoE模型，通过MoE模型中的路由器计算每个token分别选择3个专家子模型的得分，确定token1选择专家子模型a的分值为Sa1，token1选择专家子模型b的分值为Sb1，token1选择专家子模型c的分值为Sc1；token2选择专家子模型a的分值为Sa2...tokenN选择专家子模型c的分值为ScN。此后可以结合上述公式（1）和（2），计算出token1相对于所有专家子模型的归一化熵值H1...tokenN相对于所有专家子模型的归一化熵值HN，以便后续结合每个token对应的归一化熵值完成不同类型token的区分，用于完成对MoE模型的微调。

10 综上，通过计算熵值的方式体现每个分词单元选择专家子模型的确定性，可以实现筛选不确定性的token，进而便于微调阶段将其广播至所有专家子模型进行相关知识学习，使得模型具有更强的鲁棒性。

步骤S204，根据所述目标熵值将所述多个分词单元划分为全局分词单元和局部分词单元。

具体的，在上述得到每个分词单元对应的目标熵值后，考虑到目标熵值用于体现每个分词单元选择专家子模型的确定性，因此可以根据每个分词单元对应的目标熵值进行分词单元类型划分，以选择不确定性高的分词单元作为全局分词单元，以及选择不确定性低的分词单元作为局部分词单元，进而便于后续对混合专家模型进行微调。

其中，全局分词单元具体是指多个分词单元中不确定性更强的分词单元，即该类分词单元在经过路由器选择其匹配的专家子模型时，不确定性更强。相应的，局部分词单元具体是指多个分词单元中能够精准选择其匹配专家子模型的分词单元，该类分词单元不确定较弱。

20 进一步的，在上述得到每个分词单元对应的目标熵值后，在划分局部分词单元和全局分词单元时，实则是按照熵值划分分词单元的确定性和不确定性。具体实施时，可以通过如下两种方式实现：

（1）将每个分词单元对应的目标熵值与预设熵值阈值进行比较，根据比较结果选择大于所述预设熵值阈值的分词单元作为全局分词单元，以及选择小于等于所述预设熵值阈值的分词单元作为局部分词单元。

25 （2）根据每个分词单元对应的目标熵值对所述多个分词单元进行排序，获得分词单元序列，按照预设占比信息在所述分词单元序列中选择全局分词单元，并将除所述全局分词单元外的剩余分词单元作为局部分词单元。

具体的，预设熵值阈值具体是指筛选不确定分词单元和确定分词单元的阈值，其可以根据实际需求进行设定，本实施例在此不做任何限定。相应的，分词单元序列具体是指对多个分词单元按照目标熵值大小进行排序得到的序列，预设占比信息具体是指选择分词单元数量的占比值，其可以是具体数量，也可以是百分之值，本实施例在此不做任何限定。

30 基于此，在得到能够体现分词单元选择专家子模型确定性的目标熵值后，一方面，可以将每个分词单元对应的目标熵值与预设熵值阈值进行比较，根据比较结果选择大于预设熵值阈值的分词单元作为全局分词单元，以及选择小于等于预设熵值阈值的分词单元作为局部分词单元，实现后续可以结合全局分

词单元和局部分词单元完成混合专家模型的微调。

另一方面，还可以根据每个分词单元对应的目标熵值对多个分词单元进行排序，获得分词单元序列；在此基础上，可以按照预设占比信息在分词单元序列中选择全局分词单元，并且在选择的过程中，可以按照序列中分词单元的排列顺序完成，最后再将除全局分词单元外的剩余分词单元作为局部分词单元即可。

具体实施时，对多个分词单元进行排序时，可以按照目标熵值从高到低的顺序排序以组成分词单元序列，在此基础上，选择全局分词单元，即可按照从高到低的顺序选择对应预设占比信息的分词单元作为全局分词单元。

实际应用中，在按照目标熵值对分词单元进行划分时，可以根据实际需求选择适配当前场景的处理完成，本实施例在此不做任何限定。

沿用上例，在得到N个token分别对应的归一化熵值后，进一步的，可以按照归一化熵值大小对N个token进行从高到低排序，之后选择熵值前5%的token作为不确定token，不确定token包括{token1, token2, token4...}，之后将N个token除不确定token外的其他token作为确定token，确定token包括{token3, token5, token6...}。后续即可结合不确定token和确定token完成对MoE模型的微调。

综上，通过选择不同的方式完成全局分词单元和局部分词单元的划分，可以保证后续在模型微调阶段，使得模型针对不同类型的分词单元选择不同的策略完成优化，进而使得模型具备更强的预测能力。

步骤S206，将所述全局分词单元广播至所述混合专家模型中的全部专家子模型进行处理，以及将所述局部分词单元广播至所述混合专家模型中匹配所述局部分词单元的专家子模型进行处理。

具体的，在上述确定全局分词单元和局部分词单元后，进一步的，为了能够解决路由器针对分词单元选择专家子模型时选择不佳而导致预测结果不准确的问题，在模型微调阶段，可以将全局分词单元广播至混合专家模型中的全部专家子模型进行相关知识的学习，使得混合专家模型中的全部专家子模型都可以学习全局分词单元的相关知识，进而保证在应用阶段，即使遇到无法准确匹配专家子模型的分词单元时，也可以让被选择的专家子模型给出准确的预测结果。同时可以将局部分词单元广播至混合专家模型中匹配局部分词单元的专家子模型进行处理，以达到对匹配分词单元的专家子模型进行正常微调。

其中，匹配局部分词单元的专家子模型具体是指通过路由器对局部分词单元进行专家子模型选择分值计算后，可以准确确定对局部分词单元进行处理的专家子模型。

进一步的，在将局部分词单元广播给其匹配的专家子模型时，可以通过计算选择分值的方式实现。本实施例中，具体实现方式如下：

计算所述局部分词单元选择所述混合专家模型中各个专家子模型的目标选择分值；根据所述目标选择分值在所述混合专家模型中确定匹配所述局部分词单元的局部专家子模型，并将所述局部分词单元广播至所述局部专家子模型进行处理。

具体的，目标选择分值具体是指利用混合专家模型中的路由器计算局部分词单元选择各个专家子模型的分值，分值越高表明被选择的概率越高。相应的，局部专家子模型具体是指匹配局部分词单元的专家子模型，用于对局部分词单元进行处理。

基于此，在针对局部分词单元选择专家子模型时，可以采用标准 MoE 方法实现。也就是说，可以先计算局部分词单元选择混合专家模型中各个专家子模型的目标选择分值；此后可以根据目标选择分值在混合专家模型中确定匹配局部分词单元的局部专家子模型，并将局部分词单元广播至局部专家子模型进行处理即可。

5 沿用上例，在得到不确定 token 包括{token1,token2,token4...}，以及确定 token 包括{token3,token5,token6...}后，可以将不确定 token 采用广播策略，将其发送给 MoE 模型中的全部专家子模型 a、b、c 进行学习，而确定 token，则采用标准 MoE 方法，选择各个 token 匹配的专家子模型进行学习，选择时可以按照上述计算得到的分值完成。以便于后续可以整合所有专家子模型的学习结果，对模型参数进行更新。

10 综上，通过采用全局分词单元广播至全部专家子模型，而局部分词单元广播至匹配的专家子模型进行处理，可以使得全部专家子模型学习不确定性的相关知识，进而是混合专家模型具备更精准的预测能力。

步骤 S208，根据处理结果对所述混合专家模型进行调参，直至获得满足优化停止条件的目标混合专家模型。

15 具体的，在上述利用全部专家子模型对全局分词单元进行处理，以及利用匹配局部分词单元的专家子模型对局部分词单元功能处理，在此基础上，即可根据处理结果对所述混合专家模型进行调参，以获得满足优化停止条件的目标混合专家模型。

进一步的，在对混合专家模型进行调参时，可以通过比对优化停止条件的方式确定满足需求的目标混合专家模型。本实施例中，具体实现方式如下：

20 根据处理结果对所述混合专家模型进行调参，并检测调参后的混合专家模型是否满足优化停止条件；若否，选择多个候选分词单元作为所述多个分词单元，并执行将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值步骤；若是，将调参后的混合专家模型作为目标混合专家模型。

25 基于此，在对混合专家模型进行调参时，可以在每个调参阶段检测混合专家模型是否满足优化停止条件；若否，说明该模型还需要继续进行训练，因此可以选择多个候选分词单元作为多个分词单元，并执行将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值步骤；直至某个调参阶段，调参后的模型满足优化停止条件的情况下，即可将调参后的混合专家模型作为目标混合专家模型，以部署在业务场景中应用。

30 沿用上例，在全部专家子模型对不确定 token 进行处理，以及确定 token 被匹配的专家子模型进行处理后，可以合并所有专家子模型 a、b 和 c 的学习结果，以此为基础对 MoE 模型进行参数更新。当参数更新后的 MoE 模型满足业务部署需求后，即可停止微调，并将其部署在业务场景中即可。当接收到客户端提交的待分析情感文本后，可以将其输入至微调好的 MoE 模型进行处理，进而确定待分析情感文本对应的情感信息，以便于下游业务使用。

本实施例提供的模型优化方法，为了能够降低路由器选择不确定性，提高模型鲁棒性，从而保证预

测准确性，可以在模型优化阶段，通过将多个分词单元输入至混合专家模型后，先利用混合专家模型中的路由器计算每个分词单元对应的目标熵值；在此基础上，可以根据目标熵值将多个分词单元划分为全局分词单元和局部分词单元；而后可以将全局分词单元广播至混合专家模型中的全部专家子模型进行处理，同时将局部分词单元广播至混合专家模型中匹配局部分词单元的专家子模型进行处理；实现在模型优化阶段，可以将不确定分词单元广播给全部专家子模型进行学习，而确定分词单元依旧用于匹配专家子模型的学习，从而实现根据处理结果对混合专家模型进行调参，以获得满足优化停止条件的目标混合专家模型。实现在混合专家模型的微调过程中，让全部专家子模型学习不确定分词单元的相关知识，以保证在应用阶段，专家子模型具有更好的鲁棒性。

参见图 3，图 3 示出了根据本说明书一个实施例提供的一种信息处理方法的流程图，具体包括以下步骤。

步骤 S302，获取客户端提交的问题信息，根据所述问题信息构建多个问题分词单元并输入至目标混合专家模型，其中，所述目标混合专家模型通过上述方法获得；

步骤 S304，在所述目标混合专家模型中选择每个问题分词单元匹配的目标专家子模型对所述多个问题分词单元进行处理；

步骤 S306，将每个问题分词单元匹配的目标专家子模型的预测结果进行合并，获得答案信息，并将所述答案信息反馈至所述客户端。

本实施例提供的信息处理方法，关于目标混合专家模型的训练均可参见上述实施例中相同或相应的描述内容，本实施例在此不做过多赘述。

下述结合附图 4，以本说明书提供的信息处理方法在文本情感分析场景中的应用为例，对所述信息处理方法进行进一步说明。其中，图 4 示出了本说明书一个实施例提供的一种信息处理方法的处理流程图，具体包括以下步骤。

步骤 S402，将多个分词单元输入至混合专家模型，通过混合专家模型中的路由器计算每个分词单元对应的目标熵值。

其中，多个分词单元中的任意一个分词单元对应的目标熵值的计算，包括：通过混合专家模型中的路由器，计算分词单元选择混合专家模型中每个专家子模型的选择分值；确定每个专家子模型对应的选择分值的初始熵值，并通过对初始熵值进行归一化处理，获得分词单元对应的目标熵值。

步骤 S404，将每个分词单元对应的目标熵值与预设熵值阈值进行比较，根据比较结果选择大于预设熵值阈值的分词单元作为全局分词单元。

步骤 S406，根据比较结果选择小于等于预设熵值阈值的分词单元作为局部分词单元。

步骤 S408，将全局分词单元广播至混合专家模型中的全部专家子模型进行处理。

步骤 S410，计算局部分词单元选择混合专家模型中各个专家子模型的目标选择分值。

步骤 S412，根据目标选择分值在混合专家模型中确定匹配局部分词单元的局部专家子模型，并将局部分词单元广播至局部专家子模型进行处理。

步骤 S414，根据处理结果对混合专家模型进行调参，直至获得满足优化停止条件的目标混合专家模

型。

步骤 S416, 在接收到客户端提交的问题信息的情况下, 根据问题信息构建多个问题分词单元并输入至目标混合专家模型。

步骤 S418, 在目标混合专家模型中选择每个问题分词单元匹配的目标专家子模型对多个问题分词单元进行处理。

步骤 S420, 将每个问题分词单元匹配的目标专家子模型的预测结果进行合并, 获得答案信息, 并将答案信息反馈至客户端。

综上所述, 为了能够降低路由器选择不确定性, 提高模型鲁棒性, 从而保证预测准确性, 可以在模型优化阶段, 通过将多个分词单元输入至混合专家模型后, 先利用混合专家模型中的路由器计算每个分词单元对应的目标熵值; 在此基础上, 可以根据目标熵值将多个分词单元划分为全局分词单元和局部分词单元; 而后可以将全局分词单元广播至混合专家模型中的全部专家子模型进行处理, 同时将局部分词单元广播至混合专家模型中匹配局部分词单元的专家子模型进行处理; 实现在模型优化阶段, 可以将不确定分词单元广播给全部专家子模型进行学习, 而确定分词单元依旧用于匹配专家子模型的学习, 从而实现根据处理结果对混合专家模型进行调参, 以获得满足优化停止条件的目标混合专家模型。实现在混合专家模型的微调过程中, 让全部专家子模型学习不确定分词单元的相关知识, 以保证在应用阶段, 专家子模型具有更好的鲁棒性。

与上述方法实施例相对应, 本说明书还提供了模型优化装置实施例, 图 5 示出了本说明书一个实施例提供的一种模型优化装置的结构示意图。如图 5 所示, 该装置包括:

计算模块 502, 被配置为将多个分词单元输入至混合专家模型, 通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值;

划分模块 504, 被配置为根据所述目标熵值将所述多个分词单元划分为全局分词单元和局部分词单元;

处理模块 506, 被配置为将所述全局分词单元广播至所述混合专家模型中的全部专家子模型进行处理, 以及将所述局部分词单元广播至所述混合专家模型中匹配所述局部分词单元的专家子模型进行处理;

调参模块 508, 被配置为根据处理结果对所述混合专家模型进行调参, 直至获得满足优化停止条件的目标混合专家模型。

一个可选的实施例中, 所述多个分词单元中的任意一个分词单元对应的目标熵值的计算, 包括: 通过所述混合专家模型中的路由器, 计算分词单元选择所述混合专家模型中每个专家子模型的选择分值; 确定每个专家子模型对应的选择分值的初始熵值, 并通过对所述初始熵值进行归一化处理, 获得所述分词单元对应的目标熵值。

一个可选的实施例中, 所述划分模块 504 进一步被配置为: 将每个分词单元对应的目标熵值与预设熵值阈值进行比较, 根据比较结果选择大于所述预设熵值阈值的分词单元作为全局分词单元, 以及选择小于等于所述预设熵值阈值的分词单元作为局部分词单元; 或者, 根据每个分词单元对应的目标熵值对所述多个分词单元进行排序, 获得分词单元序列, 按照预设

占比信息在所述分词单元序列中选择全局分词单元，并将除所述全局分词单元外的剩余分词单元作为局部分词单元。

一个可选的实施例中，所述处理模块 506 进一步被配置为：

计算所述局部分词单元选择所述混合专家模型中各个专家子模型的目标选择分值；根据所述目标选择分值在所述混合专家模型中确定匹配所述局部分词单元的局部专家子模型，并将所述局部分词单元广播至所述局部专家子模型进行处理。

一个可选的实施例中，所述调参模块 508 进一步被配置为：

根据处理结果对所述混合专家模型进行调参，并检测调参后的混合专家模型是否满足优化停止条件；若否，选择多个候选分词单元作为所述多个分词单元，并执行将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值步骤；若是，将调参后的混合专家模型作为目标混合专家模型。

一个可选的实施例中，所述装置还包括：

分词模块，被配置为获取关联所述混合专家模型的原始样本数据，对所述原始样本数据进行预处理，获得多个初始分词单元；对所述多个初始分词单元分别进行嵌入处理，获得多个分词单元。

本实施例提供的模型优化装置，为了能够降低路由器选择不确定性，提高模型鲁棒性，从而保证预测准确性，可以在模型优化阶段，通过将多个分词单元输入至混合专家模型后，先利用混合专家模型中的路由器计算每个分词单元对应的目标熵值；在此基础上，可以根据目标熵值将多个分词单元划分为全局分词单元和局部分词单元；而后可以将全局分词单元广播至混合专家模型中的全部专家子模型进行处理，同时将局部分词单元广播至混合专家模型中匹配局部分词单元的专家子模型进行处理；实现在模型优化阶段，可以将不确定分词单元广播给全部专家子模型进行学习，而确定分词单元依旧用于匹配专家子模型的学习，从而实现根据处理结果对混合专家模型进行调参，以获得满足优化停止条件的目标混合专家模型。实现在混合专家模型的微调过程中，让全部专家子模型学习不确定分词单元的相关知识，以保证在应用阶段，专家子模型具有更好的鲁棒性。

上述为本实施例的一种模型优化装置的示意性方案。需要说明的是，该模型优化装置的技术方案与上述的模型优化方法的技术方案属于同一构思，模型优化装置的技术方案未详细描述的细节内容，均可以参见上述模型优化方法的技术方案的描述。

与上述方法实施例相对应，本说明书还提供了信息处理装置实施例，图 6 示出了本说明书一个实施例提供的一种信息处理装置的结构示意图。如图 6 所示，该装置包括：

获取信息模块 602，被配置为获取客户端提交的问题信息，根据所述问题信息构建多个问题分词单元并输入至目标混合专家模型，其中，所述目标混合专家模型通过上述方法获得；

处理分词模块 604，被配置为在所述目标混合专家模型中选择每个问题分词单元匹配的目标专家子模型对所述多个问题分词单元进行处理；

反馈信息模块 606，被配置为将每个问题分词单元匹配的目标专家子模型的预测结果进行合并，获得答案信息，并将所述答案信息反馈至所述客户端。

上述为本实施例的一种信息处理装置的示意性方案。需要说明的是，该信息处理装置的技术方案与上述的信息处理方法的技术方案属于同一构思，信息处理装置的技术方案未详细描述的细节内容，均可以参见上述信息处理方法的技术方案的描述。

图 7 示出了根据本说明书一个实施例提供的一种计算设备 700 的结构框图。该计算设备 700 的部件包括但不限于存储器 710 和处理器 720。处理器 720 与存储器 710 通过总线 730 相连接，数据库 750 用于保存数据。

计算设备 700 还包括接入设备 740，接入设备 740 使得计算设备 700 能够经由一个或多个网络 760 通信。这些网络的示例包括公用交换电话网 (PSTN, Public Switched Telephone Network)、局域网 (LAN, Local Area Network)、广域网 (WAN, Wide Area Network)、个域网 (PAN, Personal Area Network) 或诸如因特网的通信网络的组合。接入设备 740 可以包括有线或无线的任何类型的网络接口 (例如，网络接口卡 (NIC, network interface controller)) 中的一个或多个，诸如 IEEE802.11 无线局域网 (WLAN, Wireless Local Area Network) 无线接口、全球微波互联接入 (Wi-MAX, Worldwide Interoperability for Microwave Access) 接口、以太网接口、通用串行总线 (USB, Universal Serial Bus) 接口、蜂窝网络接口、蓝牙接口、近场通信 (NFC, Near Field Communication)。

在本说明书的一个实施例中，计算设备 700 的上述部件以及图 7 中未示出的其他部件也可以彼此相连接，例如通过总线。应当理解，图 7 所示的计算设备结构框图仅仅是出于示例的目的，而不是对本说明书范围的限制。本领域技术人员可以根据需要，增添或替换其他部件。

计算设备 700 可以是任何类型的静止或移动计算设备，包括移动计算机或移动计算设备 (例如，平板电脑计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等)、移动电话 (例如，智能手机)、可佩戴的计算设备 (例如，智能手表、智能眼镜等) 或其他类型的移动设备，或者诸如台式计算机或个人计算机 (PC, Personal Computer) 的静止计算设备。计算设备 700 还可以是移动式或静止式的服务器。

其中，处理器 720 用于执行如下计算机可执行指令，该计算机可执行指令被处理器执行时实现上述模型优化方法或信息处理方法的步骤。

上述为本实施例的一种计算设备的示意性方案。需要说明的是，该计算设备的技术方案与上述的模型优化方法或信息处理方法的技术方案属于同一构思，计算设备的技术方案未详细描述的细节内容，均可以参见上述模型优化方法或信息处理方法的技术方案的描述。

本说明书一实施例还提供一种计算机可读存储介质，其存储有计算机可执行指令，该计算机可执行指令被处理器执行时实现上述模型优化方法或信息处理方法的步骤。

上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是，该存储介质的技术方案与上述的模型优化方法或信息处理方法的技术方案属于同一构思，存储介质的技术方案未详细描述的细节内容，均可以参见上述模型优化方法或信息处理方法的技术方案的描述。

本说明书一实施例还提供一种计算机程序产品，包括计算机程序或指令，该计算机程序或指令被处理器执行时实现上述模型优化方法或信息处理方法的步骤。

上述为本实施例的一种计算机程序产品的示意性方案。需要说明的是，该计算机程序产品的技术方

案与上述的模型优化方法或信息处理方法的技术方案属于同一构思，计算机程序产品的技术方案未详细描述的细节内容，均可以参见上述模型优化方法或信息处理方法的技术方案的描述。

上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

所述计算机指令包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器 (ROM, Read-Only Memory)、随机存取存储器 (RAM, Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读介质包含的内容可以根据专利实践的要求进行适当的增减，例如在某些地区，根据专利实践，计算机可读介质不包括电载波信号和电信信号。

需要说明的是，对于前述的各方法实施例，为了简便描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本说明书实施例并不受所描述的动作顺序的限制，因为依据本说明书实施例，某些步骤可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本说明书实施例所必须的。

在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详述的部分，可以参见其它实施例的相关描述。

以上公开的本说明书优选实施例只是用于帮助阐述本说明书。可选实施例并没有详尽叙述所有的细节，也不限制该发明仅为所述的具体实施方式。显然，根据本说明书实施例的内容，可作很多的修改和变化。本说明书选取并具体描述这些实施例，是为了更好地解释本说明书实施例的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本说明书。

权 利 要 求

1.一种模型优化方法，其特征在于，包括：

将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值；

5 根据所述目标熵值将所述多个分词单元划分为全局分词单元和局部分词单元；

将所述全局分词单元广播至所述混合专家模型中的全部专家子模型进行处理，以及将所述局部分词单元广播至所述混合专家模型中匹配所述局部分词单元的专家子模型进行处理；

根据处理结果对所述混合专家模型进行调参，直至获得满足优化停止条件的目标混合专家模型。

2.根据权利要求1所述的模型优化方法，其特征在于，所述多个分词单元中的任意一个分词单元对
10 应的目标熵值的计算，包括：

通过所述混合专家模型中的路由器，计算分词单元选择所述混合专家模型中每个专家子模型的选择分值；

确定每个专家子模型对应的选择分值的初始熵值，并通过对所述初始熵值进行归一化处理，获得所述分词单元对应的目标熵值。

3.根据权利要求1所述的模型优化方法，其特征在于，所述根据所述目标熵值将所述多个分词单元划分为全局分词单元和局部分词单元，包括：

将每个分词单元对应的目标熵值与预设熵值阈值进行比较，根据比较结果选择大于所述预设熵值阈值的分词单元作为全局分词单元，以及选择小于等于所述预设熵值阈值的分词单元作为局部分词单元；或者，

20 根据每个分词单元对应的目标熵值对所述多个分词单元进行排序，获得分词单元序列，按照预设占比信息在所述分词单元序列中选择全局分词单元，并将除所述全局分词单元外的剩余分词单元作为局部分词单元。

4.根据权利要求1所述的模型优化方法，其特征在于，所述将所述局部分词单元广播至所述混合专家模型中匹配所述局部分词单元的专家子模型进行处理，包括：

25 计算所述局部分词单元选择所述混合专家模型中各个专家子模型的目标选择分值；

根据所述目标选择分值在所述混合专家模型中确定匹配所述局部分词单元的局部专家子模型，并将所述局部分词单元广播至所述局部专家子模型进行处理。

5.根据权利要求1所述的模型优化方法，其特征在于，所述根据处理结果对所述混合专家模型进行调参，直至获得满足优化停止条件的目标混合专家模型，包括：

30 根据处理结果对所述混合专家模型进行调参，并检测调参后的混合专家模型是否满足优化停止条件；

若否，选择多个候选分词单元作为所述多个分词单元，并执行将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值步骤；

若是，将调参后的混合专家模型作为目标混合专家模型。

6.根据权利要求1所述的模型优化方法，其特征在于，所述将多个分词单元输入至混合专家模型步骤
35 骤执行之前，还包括：

获取关联所述混合专家模型的原始样本数据，对所述原始样本数据进行预处理，获得多个初始分词单元；

对所述多个初始分词单元分别进行嵌入处理，获得多个分词单元。

7.一种信息处理方法，其特征在于，包括：

5 获取客户端提交的问题信息，根据所述问题信息构建多个问题分词单元并输入至目标混合专家模型，其中，所述目标混合专家模型通过权利要求 1 至 7 任一项所述方法获得；

在所述目标混合专家模型中选择每个问题分词单元匹配的目标专家子模型对所述多个问题分词单元进行处理；

10 将每个问题分词单元匹配的目标专家子模型的预测结果进行合并，获得答案信息，并将所述答案信息反馈至所述客户端。

8.一种模型优化装置，其特征在于，包括：

计算模块，被配置为将多个分词单元输入至混合专家模型，通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值；

划分模块，被配置为根据所述目标熵值将所述多个分词单元划分为全局分词单元和局部分词单元；

15 处理模块，被配置为将所述全局分词单元广播至所述混合专家模型中的全部专家子模型进行处理，以及将所述局部分词单元广播至所述混合专家模型中匹配所述局部分词单元的专家子模型进行处理；

调参模块，被配置为根据处理结果对所述混合专家模型进行调参，直至获得满足优化停止条件的目标混合专家模型。

9.一种信息处理装置，其特征在于，包括：

20 获取信息模块，被配置为获取客户端提交的问题信息，根据所述问题信息构建多个问题分词单元并输入至目标混合专家模型，其中，所述目标混合专家模型通过权利要求 1 至 7 任一项所述方法获得；

处理分词模块，被配置为在所述目标混合专家模型中选择每个问题分词单元匹配的目标专家子模型对所述多个问题分词单元进行处理；

25 反馈信息模块，被配置为将每个问题分词单元匹配的目标专家子模型的预测结果进行合并，获得答案信息，并将所述答案信息反馈至所述客户端。

10.一种计算设备，其特征在于，包括：

存储器和处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指令，该计算机可执行指令被处理器执行时实现权利要求 1 至 7 任意一项所述方法的步骤。

30 11.一种计算机可读存储介质，其特征在于，其存储有计算机可执行指令，该计算机可执行指令被处理器执行时实现权利要求 1 至 7 任意一项所述方法的步骤。

12.一种计算机程序产品，其特征在于，包括计算机程序或指令，该计算机程序或指令被处理器执行时实现权利要求 1 至 7 任意一项所述方法的步骤。