

训练文本预测模型的方法、文本预测方法及装置

交叉援引

本公开要求于 2023 年 04 月 23 日提交中国专利局、优先权号为 202310459343.3、发明名称为“训练文本预测模型的方法、文本预测方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

本公开涉及自然语言处理技术领域，特别是涉及一种训练文本预测模型的方法、文本预测方法及装置。

背景技术

自然语言处理（Natural Language Processing，简称为 NLP）研究的目的是让机器能够理解人类语言。其中大型语言模型（Large Language Model，简称为 LLM）是自然语言处理领域中的一个核心工具，指的是具有大规模参数（通常数以亿计或更多）的深度学习模型。

LLM 因其具有极高的学习能力而被广泛地应用于文本预测领域。其中，情景学习（In Context Learning）是目前 LLM 采用的其中一种文本预测方式。所谓情景学习指的是，给定标注数据后，LLM 进行观察和归纳，对无标签数据进行预测。由于情景学习通常没有训练过程，应用于文本预测时，预测效果不佳。因此亟需一种方式能够提高情景学习场景下基于 LLM 的文本预测效果。

发明内容

有鉴于此，本公开提供了一种训练文本预测模型的方法、文本预测方法及装置，以便于提高情景学习场景下基于 LLM 的文本预测效果。

本公开提供了如下方案：

第一方面，提供了一种训练文本预测模型的方法，所述方法包括：

获取训练数据集，所述训练数据集包括输入文本样本以及该输入文本样本对应的输出标签；

将包含输入文本样本和该输入文本样本对应的输出标签的文本序列作为文本预测模型的输入，训练所述文本预测模型；其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵

进行拼接得到第三键矩阵，以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用所述第一键矩阵对所述第二键矩阵进行更新，将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用所述第一值矩阵对所述第二值矩阵进行更新，将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

根据本公开实施例中一可实现的方式，所述大型语言模型还包括嵌入网络，所述文本预测模型还包括预测网络；

所述嵌入网络用以对所述文本序列进行嵌入处理；

若当前层 Transformer 网络为第一层 Transformer 网络，则所述上一层网络输出的特征表示为所述嵌入网络输出的特征表示；否则，所述上一层网络输出的特征表示为上一层 Transformer 网络输出的特征表示；

所述预测网络用以利用最后一层 Transformer 网络输出的特征表示预测输入文本样本对应的输出标签。

根据本公开实施例中一可实现的方式，利用所述第一键矩阵对所述第二键矩阵进行更新包括：利用所述第一键矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二键矩阵采用动量梯度下降的方式进行更新；

利用所述第一值矩阵对所述第二值矩阵进行更新包括：利用所述第一值矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二值矩阵采用动量梯度下降的方式进行更新。

根据本公开实施例中一可实现的方式，利用所述第一键矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二键矩阵采用动量梯度下降的方式进行更新包括：利用所述第一键矩阵和所述当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行逐元素求差，得到键矩阵梯度；利用所述当前层 Transformer 网络在上一轮迭代得到的键矩阵动量和所述键矩阵梯度进行加权求和，得到所述当前层 Transformer 网络在当前轮迭代得到的键矩阵动量；利用所述当前层 Transformer 网络在当前轮迭代得到的键矩阵动量和当前层 Transformer 网络在上一轮迭代得到的第二键矩阵，得到所述更新后的第二键矩阵；

利用所述第一值矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二值矩阵采用动量梯度下降的方式进行更新包括：利用所述第一值矩阵和所述当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行逐元素求差，得到值矩阵梯度；利用所述当前层 Transformer 网络在上一轮迭代得到的值矩阵动量和所述值矩阵梯度进行加权求和，得到所述当前层 Transformer 网络在当前轮迭代得到的值矩阵动量；利用所述当前层 Transformer 网络在当前轮迭代得到的值矩阵动量和当前层

Transformer 网络在上一轮迭代得到的第二值矩阵，得到所述更新后的第二值矩阵。

根据本公开实施例中一可实现的方式，每一轮迭代完成后，若确定达到预设的训练结束条件，则将各层 Transformer 网络当前迭代得到的第二键矩阵和第二值矩阵分别作为训练得到的各层 Transformer 网络的第二键矩阵和第二值矩阵进行存储。

第二方面，提供了一种文本预测方法，所述方法包括：

获取输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的输出标签；所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

根据本公开实施例中一可实现的方式，所述大型语言模型还包括嵌入网络，所述文本预测模型还包括预测网络；

所述嵌入网络用以对所述文本序列进行嵌入处理；

若当前层 Transformer 网络为第一层 Transformer 网络，则所述上一层网络输出的特征表示为所述嵌入网络输出的特征表示；否则，所述上一层网络输出的特征表示为上一层 Transformer 网络输出的特征表示；

所述预测网络用以利用最后一层 Transformer 网络输出的特征表示预测输入文本对应的输出标签。

第三方面，提供了一种训练文本预测模型的方法，所述方法包括：

获取训练数据集，所述训练数据集包括输入文本样本以及该输入文本样本对应的情感类别标签；

将包含输入文本样本和该输入文本样本对应的情感类别标签的文本序列作为文本预测模型的输入，训练所述文本预测模型；其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的 Transformer 网络；

在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特

征表示；利用所述第一键矩阵对所述第二键矩阵进行更新，将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用所述第一值矩阵对所述第二值矩阵进行更新，将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

第四方面，提供了一种情感分析方法，所述方法包括：

获取输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的情感类别标签；所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

第五方面，提供了一种文本预测方法，由云端服务器执行，所述方法包括：

获取来自用户终端的输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的输出标签；

基于所述输出标签确定对应的服务内容，将所述服务内容发送至所述用户终端；

其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

第六方面，提供了一种训练文本预测模型的装置，所述装置包括：

样本获取单元，设置为获取训练数据集，所述训练数据集包括输入文本样本以及该输入文本样本对应的输出标签样本；

模型训练单元，设置为将包含输入文本样本和该输入文本样本对应的输出标签的文本序列作为文本预测模型的输入，训练所述文本预测模型；其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代

中,利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵;将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵,以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵;利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理,得到当前层 Transformer 网络输出的特征表示;利用所述第一键矩阵对所述第二键矩阵进行更新,将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵;利用所述第一值矩阵对所述第二值矩阵进行更新,将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

第七方面,提供了一种文本预测装置,所述装置包括:

文本获取单元,设置为获取输入文本;

文本预测单元,设置为将包含所述输入文本的文本序列输入文本预测模型,获取所述文本预测模型预测得到的所述输入文本对应的输出标签;所述文本预测模型采用大型语言模型,所述大型语言模型包括多层串连的转换 Transformer 网络;其中,各 Transformer 网络分别作为当前层 Transformer 网络执行:利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵;将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵,以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵;利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理,得到当前层 Transformer 网络输出的特征表示。

根据第八方面,提供了一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现上述第一方面中任一项所述的方法的步骤。

根据第九方面,提供了一种电子设备,包括:

一个或多个处理器;以及

与所述一个或多个处理器关联的存储器,所述存储器设置为存储程序指令,所述程序指令在被所述一个或多个处理器读取执行时,执行上述第一方面中任一项所述的方法的步骤。

根据本公开提供的具体实施例,本公开公开了以下技术效果:

1) 本公开利用训练样本集对 LLM 进行训练来得到文本预测模型,这种方式实质上利用了已标注的样本训练 LLM,在第二键矩阵和第二值矩阵的更新过程中,利用了上一轮迭代得到的第二键矩阵和第二值矩阵与当前输入特征矩阵产生的第一键矩阵和第一值矩阵,既保留了历史信息又保持了当前输入文本的信息,使得 LLM 能够充分对已标注的样本进行理解和学习,从而提高 LLM 的文本预测效果。

2) 本公开在 LLM 的训练过程中,每一轮迭代过程中使用的第二键矩阵和第二值

矩阵均由上一轮迭代得到，且进行更新后用于下一轮迭代使用。这种前向优化模型的方式，仅需要优化各 Transformer 网络的第二键矩阵和第二值矩阵即可，大大缩减了需要更新的模型参数，降低了模型训练的成本，提高了效率。

3) 本公开实施例中采用动量梯度下降的方式更新各 Transformer 网络的第二键矩阵和第二值矩阵，能够加快梯度下降的速度，使得迭代效率更高，并且避免陷入局部最小值。

4) 本公开在预测过程中使用了文本预测模型训练后得到的各 Transformer 网络的第二键矩阵和第二值矩阵，与普通的情景学习相比，各 Transformer 网络的第二键矩阵和第二值矩阵包含了文本预测模型对已标注的样本数据更好的观察和理解，能够显著提高预测准确性。

当然，实施本公开的任一产品并不一定需要同时达到以上所述的所有优点。

附图说明

为了更清楚地说明本公开实施例或现有技术中的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本公开的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

- 图 1 为传统基于 LLM 进行情景学习的原理示意图；
- 图 2 为本公开实施例所适用的系统架构图；
- 图 3 为本公开实施例提供的训练文本预测模型的方法的流程图；
- 图 4 为本公开实施例提供的文本预测模型的结构示意图；
- 图 5 为本公开实施例提供的一个 Transformer 网络的原理示意图；
- 图 6 为本公开实施例提供的文本预测方法流程图；
- 图 7 为本公开实施例提供的基于 LLM 进行情景学习的原理示意图；
- 图 8 为本公开实施例提供的模型训练装置的示意性框图；
- 图 9 为本公开实施例提供的文本预测装置的示意性框图；
- 图 10 为本公开实施例提供的电子设备的示意性框图。

具体实施方式

下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本公开一部分实施例，而不是全部的实施例。基于本公开中的实施例，本领域普通技术人员所获得的所有其他实施例，都属于本公开保护的范围。

在本公开实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本公开。在本公开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。

应当理解，本文中使用的术语“和/或”仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

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”或“响应于检测”。类似地，取决于语境，短语“如果确定”或“如果检测（陈述的条件或事件）”可以被解释成为“当确定时”或“响应于确定”或“当检测（陈述的条件或事件）时”或“响应于检测（陈述的条件或事件）”。

传统基于 LLM 进行的情景学习过程如图 1 中所示，将已标注的样本数据和待预测的输入文本一起输入 LLM，由 LLM 对已标注的样本数据进行观察和归纳，从而对输入文本对应的输出标签进行预测。该情景学习过程没有训练过程，文本预测效果不佳。

若采用传统反向传播的训练方式预先基于 LLM 训练文本预测模型，即利用已标注的样本数据训练 LLM，通过损失函数产生梯度进行模型参数的更新。由于 LLM 的模型参数量巨大，传统采用反向传播的训练方式会造成成本极高且效率低下的问题。

有鉴于此，本公开提出了一种全新的思路，采用前向训练的方式预先基于 LLM 训练文本预测模型。为了方便对本公开的理解，首先对本公开所基于的系统架构进行简单描述。图 2 示出了可以应用本公开实施例的示例性系统架构。如图 2 中所示，该系统架构包括采用离线方式训练文本预测模型的模型训练装置，以及在线对输入文本进行预测的文本预测装置。

其中，模型训练装置在获取训练数据集后，可以采用本公开实施例提供的方法进行模型训练，得到文本预测模型。

文本预测装置利用已经训练得到的文本预测模型，对输入文本进行预测，得到的预测结果为输入文本对应的输出标签。例如对输入文本进行预测，得到诸如情感、意图等类别。

模型训练装置和文本预测装置可以分别设置为独立的服务器，也可以设置于同一个服务器或服务器群组，还可以设置于独立的或者同一云服务器。云服务器又称为云计算服务器或云主机，是云计算服务体系中的一项主机产品，以解决传统物理主机与虚拟专用服务器（Virtual Private Server，简称为 VPS）服务中存在的管理难度大，服

务扩展性弱的缺陷。模型训练装置和文本预测装置还可以设置于具有较强计算能力的计算机终端。

另外，文本预测装置除了在线进行文本预测之外，也可以离线进行文本预测。例如对数据库中的文本分别作为输入文本进行批量的文本预测，并将预测结果存储于数据库中供后续查询或调用。

应该理解，图 2 中的模型训练装置、文本预测装置以及文本预测模型的数目仅仅是示意性的。根据实现需要，可以具有任意数目的模型训练装置、文本预测装置以及文本预测模型。

图 3 为本公开实施例提供的训练文本预测模型的方法的流程图，该方法可以由图 2 所示系统中的模型训练装置执行。如图 3 中所示，该方法可以包括以下步骤：

步骤 302：获取训练数据集，训练数据集包括输入文本样本以及该输入文本样本对应的输出标签样本。

步骤 304：将包含输入文本样本和该输入文本样本对应的输出标签的文本序列作为文本预测模型的输入，训练文本预测模型；其中，文本预测模型采用 LLM，LLM 包括多层串连的 Transformer（转换）网络。在训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用第一键矩阵对上述第二键矩阵进行更新，将更新后的第二键矩阵作为当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用第一值矩阵对上述第二值矩阵进行更新，将更新后的第二值矩阵作为当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

由上述流程可以看出，本公开利用训练样本集对 LLM 进行训练来得到文本预测模型，这种方式实质上利用了已标注的样本训练 LLM，在第二键矩阵和第二值矩阵的更新过程中，利用了上一轮迭代得到的第二键矩阵和第二值矩阵与当前输入特征矩阵产生的第一键矩阵和第一值矩阵，既保留了历史信息又保持了当前输入文本的信息，使得 LLM 能够充分对已标注的样本进行理解和学习，从而提高 LLM 的文本预测效果。

并且，在 LLM 的训练过程中，每一轮迭代过程中使用的第二键矩阵和第二值矩阵均由上一轮迭代得到，且进行更新后用于下一轮迭代使用。这种前向优化模型的方式，仅需要优化各 Transformer 网络的第二键矩阵和第二值矩阵即可，大大缩减了需要更新的模型参数，降低了模型训练的成本，提高了效率。

需要说明的是，本公开中涉及的“第一”、“第二”等限定并不具备大小、顺序和数量等方面的限制，仅仅用以在名称上加以区分。例如“第一键矩阵”、“第二键矩阵”和“第三键矩阵”用以在名称上区分三个键矩阵。再例如“第一值矩阵”、“第二值矩阵”和“第三值矩阵”用以在名称上区分三个值矩阵。

下面分别对上述流程中的各步骤进行详细描述。首先结合实施例对上述步骤 302 即“获取训练数据集”进行详细描述。

本步骤中获取的训练数据集包括多个训练样本。训练样本包括输入文本样本以及该输入文本样本对应的输出标签样本，每个训练样本可以表示为 (x_i, y_i) ，其中 x_i 为第 i 个输入文本样本， y_i 为 x_i 对应的输出标签样本，即为 x_i 标注的输出标签。

上述输入文本样本可以是句子、段落、文章、短语等等。输出标签在不同应用场景下可以对应不同的内容，可以是类别标签，也可以是文本标签。例如，在情感分析场景下，输出标签可以是情感类型标签，即对输入文本样本标注的情感类型。再例如，在意图识别场景下，输出标签可以是意图类型标签，即对输入文本样本标注的意图类型。再例如，在自动问答场景下，输出标签可以是答案文本标签，即将输入文本样本作为问题，针对问题标注的答案文本。

上述输出标签样本可以采用人工的方式进行标注，也可以采用已有的挖掘方式自动进行标注，本公开对此不做限制。

下面结合实施例重点对上述步骤 304 即“将包含输入文本样本和该输入文本样本对应的输出标签的文本序列作为文本预测模型的输入，训练文本预测模型”进行详细描述。

本公开实施例中文本预测模型基于 LLM 实现，LLM 为预训练的大型语言模型，本公开实施例在 LLM 的基础上进一步进行模型优化，以使得 LLM 能够快速理解已标注的样本集训练数据集。为了充分利用 LLM 已有的语言理解能力，在进行训练的过程中，利用各输入文本样本和该输入文本样本对应的输出标签分别构建文本序列，例如将输入文本样本 x_i 与对应的输出标签 y_i 进行拼接得到文本序列 T_i 输入文本预测模型即 LLM。在拼接 x_i 和 y_i 时，可以增加对 x_i 和 y_i 的指示信息。例如，输入文本样本“食物太好吃了”，输出标签“积极”，可以构成文本序列“评价：食物太好吃了。情感：积极”。其中“评价”和“情感”是对输入文本样本和输出标签的指示信息。

本公开实施例所采用的文本预测模型的整体框架可以如图 4 中所示，主要包括特征提取网络和预测网络。

其中特征提取网络用以从输入的文本序列中提取特征表示。预测网络用以利用特征提取网络提取的特征表示预测输入文本样本的输出标签。除了该框架之外，也可以

采用其他诸如编码器-解码器框架等，在此不做一一列举，但框架均基于 LLM 实现，例如特征提取网络基于 LLM 实现，编码器基于 LLM 实现等等，本公开实施例中仅以图 4 所示框架为例。

鉴于本公开的文本预测模型基于 LLM 实现，因此特征提取网络可以具体包括嵌入网络和多层串连的 Transformer（转换）网络。

嵌入网络用以对输入的文本序列进行 Embedding（嵌入）处理。更具体地，可以对输入的文本序列的各 Token（元素）进行 Embedding 处理。文本序列的各 Token 指的是构成文本序列的元素。对于文本序列而言，将文本序列切分为字符或者词语序列，则文本序列中的字符或者词语、以及起始符、分隔符均为 Token。

上述基于 Token 的 Embedding 处理至少包括：词 Embedding 和位置 Embedding。词 Embedding，即将各 Token 进行词向量编码，得到词向量表示。位置 Embedding，即将各 Token 在待预测文本序列中的位置进行编码，得到位置的表示。

Transformer 网络是一个采用自注意力机制对输入的各 Token 进行编码以转换为特征表示的模型。本公开实施例提供的前向学习的方式，仅需要对 Transformer 中的自注意力部分进行修改。

图 5 为本公开实施例提供的一个 Transformer 网络的原理性示意图，如图 5 中所示，每一个 Transformer 网络都包含“横向”对特征表示的处理以及“纵向”对模型参数的更新处理。需要说明的是，这里的“横向”和“纵向”是图中所示出的方向，仅为了方便描述而使用的名称，并不具有真实的方向含义。

如图 5 中的“横向”处理过程中，当前 Transformer 网络首先利用上一层网络输出的特征表示 X_t^{l-1} 确定第一键矩阵 K_X 、第一值矩阵 V_X 和第一查询矩阵 Q 。其中，若当前 Transformer 网络为第一层 Transformer 网络，则上述 X_t^{l-1} 为嵌入网络输出的特征表示。若当前 Transformer 网络并非第一层 Transformer 网络，则上述 X_t^{l-1} 为上一层 Transformer 网络输出的特征表示。图中以当前层 Transformer 网络为第 l 层 Transformer 网络为例。

图中的下标 t 表示文本预测模型训练过程中的第 t 轮迭代。每一轮迭代可以从训练数据集中采集一个 batch（批）的训练样本用来输入文本预测模型进行训练，该部分与传统模型训练类似，不做详述。

上述第一键矩阵 K_X 、第一值矩阵 V_X 和第一查询矩阵 Q_X 可以采用如下公式确定：

$$K_X = W_k X_t^{l-1} \quad (1)$$

$$V_X = W_v X_t^{l-1} \quad (2)$$

$$Q=W_Q X_t^{l-1} \quad (3)$$

其中， W_Q 、 W_k 和 W_v 为权重矩阵，是 LLM 在预训练过程中已经学习到的参数。

然后将第一键矩阵 K_X 与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵 K_{t-1}^l 进行拼接得到第三键矩阵 K ，以及将第一值矩阵 V_X 与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵 V_{t-1}^l 进行拼接得到第三值矩阵 V 。

其中， K 和 V 可以表示为：

$$K=\{K_{t-1}^l \| K_X\} \quad (4)$$

$$V=\{V_{t-1}^l \| V_X\} \quad (5)$$

其中， $\|$ 表示拼接的处理。

各 Transformer 网络的第二键矩阵和第二值矩阵是前向训练过程中需要学习的模型参数，在每一轮迭代中不断更新，每一个 Transformer 网络均对应第二键矩阵和第二值矩阵。

再利用第三键矩阵 K 、第三值矩阵 V 和第一查询矩阵 Q 进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示 X_t^l 。其中自注意力机制的处理是注意力模块和 FFN 模块的处理，在此不做详述，可以表示为：

$$X_t^l=\text{Transformer}(K,V,Q) \quad (6)$$

其中， $\text{Transformer}()$ 为注意力模块和 FFN 模块的处理函数。

上述“横向”处理过程中，将第一键矩阵 K_X 与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵 K_{t-1}^l 进行拼接，以及将第一值矩阵 V_X 与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵 V_{t-1}^l 进行拼接，可以看作是混合了“过去与现在”的信息从而得到第三键矩阵 K 、第三值矩阵 V ，符合情景学习的本质。

对于“纵向”处理过程，首先利用第一键矩阵 K_X 对第二键矩阵 K_{t-1}^l 进行更新，将更新后的第二键矩阵 K_t^l 作为当前层 Transformer 网络在当前轮迭代得到的第二键矩阵，以供下一轮迭代使用。利用第一值矩阵 V_X 对第二值矩阵 V_{t-1}^l 进行更新，将更新后的第二值矩阵 V_t^l 作为当前层 Transformer 网络在当前轮迭代得到的第二值矩阵，以供下一轮迭代使用。可以看出，该参数的更新过程是一个前向更新过程。

作为其中一种可实现的方式，上述第二键矩阵和第二值矩阵的更新可以采用动量梯度下降的方式。所谓动量梯度下降法就是在梯度下降中，每一次迭代中参数更新的方向与本次迭代采用的动量的方向相反。也可以认为每一次迭代中参数更新的方向是当前梯度的反方向与上一次参数更新方向加权组合而成。也就是说，参数更新的方向

不只取决于当前梯度的方向，还取决于过去的参数更新方向。

以更新第二键矩阵为例，可以首先利用第一键矩阵 K_X 和当前层 Transformer 网络在上一轮迭代得到的第二键矩阵 K_{t-1}^l 进行逐元素求差，得到键矩阵梯度 $G_{K_t^l}$ 。再利用当前层 Transformer 网络在上一轮迭代得到的键矩阵动量 $M_{K_{t-1}^l}$ 和键矩阵梯度 $G_{K_t^l}$ 进行加权求和，得到当前层 Transformer 网络在当前轮迭代得到的键矩阵动量 $M_{K_t^l}$ 。然后利用当前层 Transformer 网络在当前轮迭代得到的键矩阵动量 $M_{K_t^l}$ 和当前层 Transformer 网络在上一轮迭代得到的第二键矩阵 K_{t-1}^l ，得到更新后的第二键矩阵 K_t^l 。整个过程可以如下公式所示：

$$G_{K_t^l} = K_X - K_{t-1}^l \quad (7)$$

$$M_{K_t^l} = (1 - \beta) M_{K_{t-1}^l} + \beta G_{K_t^l} \quad (8)$$

$$K_t^l = K_{t-1}^l + \eta M_{K_t^l} \quad (9)$$

其中， β 和 η 分别为动量权重和更新步长，均为预设的超参数，可以采用经验值或者实验值，例如 β 取 0.9， η 取 0.01。

第二值矩阵的更新过程为：首先利用第一值矩阵 V_X 和当前层 Transformer 网络在上一轮迭代得到的第二值矩阵 V_{t-1}^l 进行逐元素求差，得到值矩阵梯度 $G_{V_t^l}$ 。再利用当前层 Transformer 网络在上一轮迭代得到的值矩阵动量 $M_{V_{t-1}^l}$ 和值矩阵梯度 $G_{V_t^l}$ 进行加权求和，得到当前层 Transformer 网络在当前轮迭代得到的值矩阵动量 $M_{V_t^l}$ 。然后利用当前层 Transformer 网络在当前轮迭代得到的值矩阵动量 $M_{V_t^l}$ 和当前层 Transformer 网络在上一轮迭代得到的第二值矩阵 V_{t-1}^l ，得到更新后的第二值矩阵 V_t^l 。整个过程可以如下公式所示：

$$G_{V_t^l} = V_X - V_{t-1}^l \quad (10)$$

$$M_{V_t^l} = (1 - \beta) M_{V_{t-1}^l} + \beta G_{V_t^l} \quad (11)$$

$$V_t^l = V_{t-1}^l + \eta M_{V_t^l} \quad (12)$$

本公开实施例中在进行模型训练时，输入为输入文本样本和输出标签，Transformer 网络通过自注意力处理来学习输入文本样本和输出标签之间的关系，上述过程中的各键矩阵和值矩阵是表达这些关系的关键载体，上述的键矩阵梯度和值矩阵梯度可以看作是两轮迭代学习到的输入文本样本和输出标签之间的关系差异，通过不断前向更新，使得得到的关系趋于一致。

上述更新过程实际上是将上一轮迭代得到的第二键矩阵和第二值矩阵与当前输入特征矩阵产生的第一键矩阵和第一值矩阵进行信息“混合”而产生更新后的第二键矩阵和第二值矩阵，即保留了历史信息又保持了当前输入文本的信息，实现了参数的平滑更新，提高了模型的文本预测能力。

从上述实施例中的可以看出，整个模型的更新过程中仅仅涉及到第二键矩阵和第二值矩阵这两个主要参数的更新，而对于模型基本网络的参数，例如嵌入网络、自注意力模块、FFN 模块和预测网络等涉及的参数均无需进行更新，大大缩减了需要更新的参数量，降低了计算成本，提高了效率。

每一轮迭代完成后，可以判断是否满足预设的训练结束条件，如果是，则停止迭代，训练结束；否则进行下一轮迭代。训练结束后，将各层 Transformer 网络当前迭代得到的第二键矩阵和第二值矩阵分别作为训练得到的各层 Transformer 网络的第二键矩阵 K_T^l 和第二值矩阵 V_T^l 进行存储。 K_T^l 和 V_T^l 可以设置为进行文本预测过程中使用。

其中，训练结束条件可以包括但不限于：在验证集上的准确率达到预设准确率阈值，迭代次数达到预设的迭代次数阈值，等等。

本公开实施例中采用动量梯度下降的方式更新各 Transformer 网络的第二键矩阵和第二值矩阵，能够加快梯度下降的速度，使得迭代效率更高，并且避免陷入局部最小值。另外，除了上述动量梯度下降的方式之外，也可以采用其他的梯度下降方式，在此不做一一列举。

图 6 为本公开实施例提供的文本预测方法流程图，该方法可以由图 2 所示系统架构中的文本预测装置执行。如图 6 中所示，该方法可以包括以下步骤：

步骤 602：获取输入文本。

本步骤中涉及的输入文本是没有标注对应输出标签的文本，可以是句子、段落、文章、短语等等。

步骤 604：将包含输入文本的文本序列输入文本预测模型，获取文本预测模型预测得到的输入文本对应的输出标签；文本预测模型采用大型语言模型，大型语言模型包括多层串连的转换 Transformer 网络。其中，各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

可以仅将输入文本作为文本序列。也可以将输入文本以及输入文本的指示信息构成文本序列。还可以将输入文本与预设的提示文本进行拼接，构成文本序列。其中提示文本可以根据实际场景的需求预先设置，例如在进行情感识别时，可以采用诸如“该

句子所表达的情感为【Mask】”。其中【Mask】为掩码部分，文本预测模型通过对文本序列的编码和预测，实际上是实现对【mask】部分内容的预测。

文本预测模型的结构可以参见之前实施例中的相关记载，若采用图 4 所示框架，则将文本序列输入嵌入网络，嵌入网络用以对输入的文本序列进行 Embedding 处理。然后经过串连的多个 Transformer 网络进行编码，得到文本序列的特征表示。最后由预测网络利用文本序列的特征表示预测输入文本对应的输出标签。

其中，每一个 Transformer 网络的处理过程可以如图 7 中所示，若上一层网络输出的特征表示为 X_q^{l-1} ，则首先利用 X_q^{l-1} 确定第一键矩阵 K_X 、第一值矩阵 V_X 和第一查询矩阵 Q 。其中，若当前 Transformer 网络为第一层 Transformer 网络，则上述 X_q^{l-1} 为嵌入网络输出的特征表示。若当前 Transformer 网络并非第一层 Transformer 网络，则上述 X_q^{l-1} 为上一层 Transformer 网络输出的特征表示。图中以当前层 Transformer 网络为第 l 层 Transformer 网络为例。

上述第一键矩阵 K_X 、第一值矩阵 V_X 和第一查询矩阵 Q_X 可以采用如下公式确定：

$$K_X = W_k X_q^{l-1} \quad (13)$$

$$V_X = W_v X_q^{l-1} \quad (14)$$

$$Q = W_Q X_q^{l-1} \quad (15)$$

然后将第一键矩阵 K_X 与预先训练得到的当前层 Transformer 网络的第二键矩阵 K_T^l 进行拼接得到第三键矩阵 K ，以及将第一值矩阵 V_X 与预先训练得到的当前层 Transformer 网络的第二值矩阵 V_T^l 进行拼接得到第三值矩阵 V 。

其中， K 和 V 可以表示为：

$$K = \{ K_T^l \| K_X \} \quad (16)$$

$$V = \{ V_T^l \| V_X \} \quad (17)$$

再利用第三键矩阵 K 、第三值矩阵 V 和第一查询矩阵 Q 进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示 X_q^l 。其中自注意力机制的处理是注意力模块和 FFN 模块的处理，在此不做详述，可以表示为：

$$X_q^l = \text{Transformer}(K, V, Q) \quad (18)$$

从上述预测过程可以看出，该预测过程使用了在模型训练后得到的 K_T^l 和 V_T^l ，与普通的情景学习相比， K_T^l 和 V_T^l 包含了文本预测模型对已标注的样本数据更好的观察和理解，能够显著提高预测准确性。

从上述训练和预测过程可以看出，本公开实施例采用的情景学习的原理如图 7 中所示，先利用训练样本集中的样本数据训练 LLM，最终得到所有 Transformer 网络的第二键矩阵 K_T 和第二值矩阵 V_T 。在预测过程中，利用已经训练得到的 LLM 对输入文本进行预测，预测过程中各 Transformer 需要使用预先训练得到的 K_T 和 V_T 进行自注意力处理，具体过程参见之前实施例中的相关记载。与图 1 所示现有的情景学习存在显著区别。

本公开实施例提供的上述方法可以应用于多种应用场景，包括但不限于：情感分析、意图识别、问答处理等。下面以情感分析为例，对上述实施例提供的方法进行举例描述。

首先获取已标注的样本数据构建训练数据集，训练数据集包括多个训练样本，各训练样本包括输入文本样本以及该输入文本样本对应的情感类别标签。例如表 1 中所示：

表 1

输入文本样本	情感类别 标签
食物太好吃了	积极
食物太难吃了	消极
糟糕的菜	消极
这道菜汤汁浓郁	积极
香气扑鼻，鲜美 无比	积极
.....	

上述表 1 中提供的情感类别是以“积极”、“中性”和“消极”等为例，也可以划分为其他更粗粒度，或者更细粒度的情感类别。例如划分为“兴奋”、“高兴”、“平和”、“伤心”、“生气”、“愤怒”等等。

将包含输入文本样本和情感类别标签的文本序列作为文本预测模型的输入，训练文本预测模型。例如，构建文本序列“评价：食物太好吃了。情感类别：积极”。

该文本预测模型是在预训练得到的 LLM 的基础上进行模型优化得到的。文本预测模型的结构和训练过程可以参见上面方法实施例中的相关记载。其中的核心内容是，LLM 包括多层串连的转换 Transformer 网络。在训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定

第一键矩阵、第一值矩阵和第一查询矩阵；将第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用第一键矩阵对上述第二键矩阵进行更新，将更新后的第二键矩阵作为当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用第一值矩阵对上述第二值矩阵进行更新，将更新后的第二值矩阵作为当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。其他细节在此不做赘述。

在进行模型训练时，Transformer 网络通过自注意力处理来学习输入文本样本和情感类别标签之间的关系，上述过程中的各键矩阵和值矩阵是表达这些关系的关键载体，上述的键矩阵梯度和值矩阵梯度可以看作是两轮迭代学习到的输入文本样本和情感类别标签之间的关系差异，通过不断前向更新，使得得到的关系趋于一致。

模型训练过程中，文本预测模型在每一轮迭代中均对历史信息进行了保留，从而对已标注的样本数据进行了充分的理解和学习。模型训练结束后，将训练得到的各层 Transformer 网络的第二键矩阵 K_T^l 和第二值矩阵 V_T^l 进行存储，用于进行情感分析过程中使用。

在利用已经训练得到的文本预测模型进行情感分析时，对于未标注的输入文本，例如“这道菜真让人爱不释口”，作为其中一种可实现的方式，可以将“评价：这道菜真让人爱不释口”作为文本序列输入文本预测模型，获取文本预测模型预测得到的输入文本对应的情感类别。

作为另一种可实现的方式，可以将输入文本与提示文本进行拼接得到的文本序列输入文本预测模型，获取文本预测模型预测得到的输入文本对应的情感类别。其中，提示文本可以采用诸如“该句子所表达的情感为【Mask】”。例如，将输入文本“这道菜汤汁浓郁”与提示文本进行拼接后，可以得到文本序列“这道菜汤汁浓郁，该句子所表达的情感为【Mask】”。其中【Mask】为掩码部分，文本预测模型通过对文本序列的编码和预测，实际上是实现对【mask】部分内容的预测，预测结果实际上是映射到情感类别空间上的具体情感类别上。

同样，文本预测模型采用大型语言模型，其结构和原理参见之前实施例中的相关记载，在此不做赘述。其核心内容是，大型语言模型包括多层串连的转换 Transformer 网络。其中，各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵 K_T^l 进行拼接得到第三键矩阵，以及将第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵 V_T^l 进行拼

接得到第三值矩阵；利用第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

作为其中一种可实现的方式，上述训练文本预测模型的方法或文本预测方法可以由云端服务器执行。云服务器又称为云计算服务器或云主机，是云计算服务体系中的一项主机产品，以解决传统物理主机与虚拟专用服务器（VPS, Virtual Private Server）服务中存在的管理难度大，服务扩展性弱的缺陷。

云端服务器获取到来自用户终端的输入文本后；将包含输入文本的文本序列输入文本预测模型，获取文本预测模型预测得到的输入文本对应的输出标签；基于输出标签确定对应的服务内容，将服务内容发送至所述用户终端。例如，获取用户在智能客服系统中输入的文本，对该文本进行情感分析。然后依据情感分析得到的情感类别向用户提供对应的服务内容。例如若识别出用户很生气，可以转接人工客服进行处理。若识别出用户很高兴，可以向用户发送推广信息，等等。

其中，所述文本预测模型的原理和结构可以参见之前方法实施例中的相关记载，在此不做赘述。

其中上述终端设备可以包括但不限于诸如：智能移动终端、智能家居设备、可穿戴式设备、PC（Personal Computer，个人计算机）等。其中智能移动设备可以包括诸如手机、平板电脑、笔记本电脑、PDA（Personal Digital Assistant，个人数字助理）、互联网汽车等。智能家居设备可以包括智能电视、智能音箱、智能冰箱等等。可穿戴式设备可以包括诸如智能手表、智能眼镜、虚拟现实设备、增强现实设备、混合现实设备（即可以支持虚拟现实和增强现实的设备）等等。

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。

根据另一方面的实施例，提供了一种训练文本预测模型的装置，该装置对应于图 2 所示系统中的模型训练装置。图 8 示出根据一个实施例的该模型训练装置的示意性框图，如图 8 所示，该装置 800 可以包括：样本获取单元 801 和模型训练单元 802。其中各单元的主要功能如下：

样本获取单元 801，设置为获取训练数据集，训练数据集包括输入文本样本以及该输入文本样本对应的输出标签样本。

模型训练单元 802，设置为将包含输入文本样本和该输入文本样本对应的输出标签的文本序列作为文本预测模型的输入，训练文本预测模型。其中，文本预测模型采

用 LLM，LLM 包括多层串连的转换 Transformer 网络；在训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用第一键矩阵对第二键矩阵进行更新，将更新后的第二键矩阵作为当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用第一值矩阵对第二值矩阵进行更新，将更新后的第二值矩阵作为当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

更进一步地，上述 LLM 还可以包括嵌入网络，上述文本预测模型还包括预测网络。

嵌入网络用以对文本序列进行嵌入处理。

若当前层 Transformer 网络为第一层 Transformer 网络，则上述的上一层网络输出的特征表示为嵌入网络输出的特征表示；否则，上述的上一层网络输出的特征表示为上一层 Transformer 网络输出的特征表示。

预测网络用以利用最后一层 Transformer 网络输出的特征表示预测输入文本样本对应的输出标签。

作为其中一种可实现的方式，在当前层 Transformer 网络中，可以利用第一键矩阵对当前层 Transformer 网络在上一轮迭代得到的第二键矩阵采用动量梯度下降的方式进行更新；利用第一值矩阵对当前层 Transformer 网络在上一轮迭代得到的第二值矩阵采用动量梯度下降的方式进行更新。

作为其中一种可实现的方式，在当前层 Transformer 网络中，可以利用第一键矩阵和当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行逐元素求差，得到键矩阵梯度；利用当前层 Transformer 网络在上一轮迭代得到的键矩阵动量和键矩阵梯度进行加权求和，得到当前层 Transformer 网络在当前轮迭代得到的键矩阵动量；利用当前层 Transformer 网络在当前轮迭代得到的键矩阵动量和当前层 Transformer 网络在上一轮迭代得到的第二键矩阵，得到更新后的第二键矩阵。

可以利用第一值矩阵和当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行逐元素求差，得到值矩阵梯度；利用当前层 Transformer 网络在上一轮迭代得到的值矩阵动量和值矩阵梯度进行加权求和，得到当前层 Transformer 网络在当前轮迭代得到的值矩阵动量；利用当前层 Transformer 网络在当前轮迭代得到的值矩阵动量和当前层 Transformer 网络在上一轮迭代得到的第二值矩阵，得到更新后的第二值矩阵。

更进一步地，每一轮迭代完成后，若确定达到预设的训练结束条件，则模型训练单元 802 将各层 Transformer 网络当前迭代得到的第二键矩阵和第二值矩阵分别作为训练得到的各层 Transformer 网络的第二键矩阵和第二值矩阵进行存储。

根据再一方面的实施例，提供了一种文本预测装置。图 9 示出根据一个实施例的文本预测装置的示意性框图，如图 9 所示，该装置 900 可以包括：文本获取单元 901 和文本预测单元 902。其中各单元的主要功能如下：

文本获取单元 901，设置为获取输入文本。

文本预测单元 902，设置为将包含输入文本的文本序列输入文本预测模型，获取文本预测模型预测得到的输入文本对应的输出标签；文本预测模型采用 LLM，LLM 包括多层串连的转换 Transformer 网络；其中，各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

更进一步地，LLM 还包括嵌入网络，上述文本预测模型还包括预测网络。

嵌入网络用以对文本序列进行嵌入处理。

若当前层 Transformer 网络为第一层 Transformer 网络，则上述的上一层网络输出的特征表示为嵌入网络输出的特征表示；否则，上述的上一层网络输出的特征表示为上一层 Transformer 网络输出的特征表示。

预测网络用以利用最后一层 Transformer 网络输出的特征表示预测输入文本对应的输出标签。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于装置实施例而言，由于其基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

需要说明的是，本公开所涉及的用户信息（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

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。本公开实施例所提供的各个功能单元可以在上述实施例中的方法或者类似的运算装置中运行，也可以作为计算机可读存储介质的一部分进行存储。

另外，本公开实施例还提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现前述方法实施例中任一项所述的方法的步骤。

可选地，上述计算机程序被处理器执行时实现如下步骤的程序代码：

获取训练数据集，所述训练数据集包括输入文本样本以及该输入文本样本对应的输出标签；

将包含输入文本样本和该输入文本样本对应的输出标签的文本序列作为文本预测模型的输入，训练所述文本预测模型；其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用所述第一键矩阵对所述第二键矩阵进行更新，将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用所述第一值矩阵对所述第二值矩阵进行更新，将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

可选地，上述计算机程序被处理器执行时实现如下步骤的程序代码：

获取输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的输出标签；所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

可选地，上述计算机程序被处理器执行时实现如下步骤的程序代码：

获取训练数据集，所述训练数据集包括输入文本样本以及该输入文本样本对应的情感类别标签；

将包含输入文本样本和该输入文本样本对应的情感类别标签的文本序列作为文本预测模型的输入，训练所述文本预测模型；其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的 Transformer 网络；

在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用所述第一键矩阵对所述第二键矩阵进行更新，将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用所述第一值矩阵对所述第二值矩阵进行更新，将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

可选地，上述计算机程序被处理器执行时实现如下步骤的程序代码：

获取输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的情感类别标签；所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

可选地，上述计算机程序被处理器执行时实现如下步骤的程序代码：

获取来自用户终端的输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的输出标签；

基于所述输出标签确定对应的服务内容，将所述服务内容发送至所述用户终端；

其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

计算机可读存储介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了可读程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读存储介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

计算机可读存储介质中包含的程序代码可以用任何适当的介质传输，包括但不限于无线、有线、光缆、射频等等，或者上述的任意合适的组合。

以及一种电子设备，包括：

一个或多个处理器；以及

与所述一个或多个处理器关联的存储器，所述存储器用于存储程序指令，所述程序指令在被所述一个或多个处理器读取执行时，执行前述方法实施例中任一项所述的方法的步骤。

本公开还提供了一种计算机程序产品，包括计算机程序，该计算机程序在被处理器执行时实现前述方法实施例中任一项所述的方法的步骤。

其中，图 10 示例性的展示出了电子设备的架构，具体可以包括处理器 1010，视频显示适配器 1011，磁盘驱动器 1012，输入/输出接口 1013，网络接口 1014，以及存储器 1020。上述处理器 1010、视频显示适配器 1011、磁盘驱动器 1012、输入/输出接口 1013、网络接口 1014，与存储器 1020 之间可以通过通信总线 1030 进行通信连接。

其中，处理器 1010 可以采用通用的 CPU、微处理器、应用专用集成电路（Application Specific Integrated Circuit，ASIC）、或者一个或多个集成电路等方式实现，设置为执行相关程序，以实现本公开所提供的技术方案。

存储器 1020 可以采用 ROM（Read Only Memory，只读存储器）、RAM（Random Access Memory，随机存取存储器）、静态存储设备、动态存储设备等形式实现。存储器 1020 可以存储设置为控制电子设备 1000 运行的操作系统 1021，设置为控制电子设备 1000 的低级别操作的基本输入输出系统（BIOS）1022。另外，还可以存储网页浏览器 1023，数据存储管理系统 1024，以及模型训练装置/文本预测装置 1025 等等。上述模型训练装置/文本预测装置 1025 就可以是本公开实施例中具体实现前述各步骤操

作的应用程序。总之，在通过软件或者固件来实现本公开所提供的技术方案时，相关的程序代码保存在存储器 1020 中，并由处理器 1010 来调用执行。

输入/输出接口 1013 设置为连接输入/输出模块，以实现信息输入及输出。输入输出/模块可以作为组件配置在设备中（图中未示出），也可以外接于设备以提供相应功能。其中输入设备可以包括键盘、鼠标、触摸屏、麦克风、各类传感器等，输出设备可以包括显示器、扬声器、振动器、指示灯等。

网络接口 1014 设置为连接通信模块（图中未示出），以实现本设备与其他设备的通信交互。其中通信模块可以通过有线方式（例如 USB、网线等）实现通信，也可以通过无线方式（例如移动网络、WIFI、蓝牙等）实现通信。

总线 1030 包括一通路，在设备的各个组件（例如处理器 1010、视频显示适配器 1011、磁盘驱动器 1012、输入/输出接口 1013、网络接口 1014，与存储器 1020）之间传输信息。

需要说明的是，尽管上述设备仅示出了处理器 1010、视频显示适配器 1011、磁盘驱动器 1012、输入/输出接口 1013、网络接口 1014，存储器 1020，总线 1030 等，但是在具体实施过程中，该设备还可以包括实现正常运行所必需的其他组件。此外，本领域的技术人员可以理解的是，上述设备中也可以仅包含实现本公开方案所必需的组件，而不必包含图中所示的全部组件。

本公开实施例，还提供了一种处理器。图 11 示出了用来实现本公开实施例的处理器结构示意图。如图 11 所示，该处理器 1100 设置为运行程序，其中，程序被该处理器运行时执行上述实施例中的方法。

在本公开实施例，上述处理器 1100 可以执行训练文本预测模型的方法的运行程序。

可选地，在本实施例中，处理器 1100 可以被设置为执行下述步骤：

获取训练数据集，所述训练数据集包括输入文本样本以及该输入文本样本对应的输出标签；

将包含输入文本样本和该输入文本样本对应的输出标签的文本序列作为文本预测模型的输入，训练所述文本预测模型；其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩

阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用所述第一键矩阵对所述第二键矩阵进行更新，将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用所述第一值矩阵对所述第二值矩阵进行更新，将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

可选地，在本实施例中，处理器 1100 可以被设置为执行下述步骤：

获取输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的输出标签；所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

可选地，在本实施例中，处理器 1100 可以被设置为执行下述步骤：

获取训练数据集，所述训练数据集包括输入文本样本以及该输入文本样本对应的情感类别标签；

将包含输入文本样本和该输入文本样本对应的情感类别标签的文本序列作为文本预测模型的输入，训练所述文本预测模型；其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的 Transformer 网络；

在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用所述第一键矩阵对所述第二键矩阵进行更新，将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用所述第一值矩阵对所述第二值矩阵进行更新，将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

可选地，在本实施例中，处理器 1100 可以被设置为执行下述步骤：

获取输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的情感类别标签；所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

可选地，在本实施例中，处理器 1100 可以被设置为执行下述步骤：

获取来自用户终端的输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的输出标签；

基于所述输出标签确定对应的服务内容，将所述服务内容发送至所述用户终端；

其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

通过以上的实施方式的描述可知，本领域的技术人员可以清楚地了解到本公开可借助软件加必需的通用硬件平台的方式来实现。基于这样的理解，本公开的技术方案本质上或者说对现有技术做出贡献的部分可以以计算机程序产品的形式体现出来，该计算机程序产品可以存储在存储介质中，如 ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本公开各个实施例或者实施例的某些部分所述的方法。

以上对本公开所提供的技术方案进行了详细介绍，本文中应用了具体个例对本公开的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本公开的方法及其核心思想；同时，对于本领域的一般技术人员，依据本公开的思想，在具体实施方式及应用范围上均会有改变之处。综上所述，本说明书内容不应理解为对本公开的

限制。

工业实用性

本公开实施例提供的方案可以应用于训练文本预测模型的过程中，获取训练数据集，训练数据集包括输入文本样本以及该输入文本样本对应的输出标签；将包含输入文本样本和该输入文本样本对应的输出标签的文本序列作为文本预测模型的输入，训练文本预测模型；其中，文本预测模型采用大型语言模型，大型语言模型包括多层串连的转换 Transformer 网络；在训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用第一键矩阵对第二键矩阵进行更新，将更新后的第二键矩阵作为当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用第一值矩阵对第二值矩阵进行更新，将更新后的第二值矩阵作为当前层 Transformer 网络在当前轮迭代得到的第二值矩阵，进而解决了训练文本预测模型的效率低的技术问题。

权利要求书

1. 一种训练文本预测模型的方法，所述方法包括：

获取训练数据集，所述训练数据集包括输入文本样本以及所述输入文本样本对应的输出标签；

将包含输入文本样本和所述输入文本样本对应的输出标签的文本序列作为文本预测模型的输入，训练所述文本预测模型；其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用所述第一键矩阵对所述第二键矩阵进行更新，将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用所述第一值矩阵对所述第二值矩阵进行更新，将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

2、根据权利要求 1 所述的方法，其中，所述大型语言模型还包括嵌入网络，所述文本预测模型还包括预测网络；

所述嵌入网络用以对所述文本序列进行嵌入处理；

若当前层 Transformer 网络为第一层 Transformer 网络，则所述上一层网络输出的特征表示为所述嵌入网络输出的特征表示；否则，所述上一层网络输出的特征表示为上一层 Transformer 网络输出的特征表示；

所述预测网络用以利用最后一层 Transformer 网络输出的特征表示预测输入文本样本对应的输出标签。

3、根据权利要求 1 所述的方法，其中，利用所述第一键矩阵对所述第二键矩阵进行更新包括：利用所述第一键矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二键矩阵采用动量梯度下降的方式进行更新；

利用所述第一值矩阵对所述第二值矩阵进行更新包括：利用所述第一值矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二值矩阵采用动量梯度下降的方式进行更新。

4、根据权利要求 3 所述的方法，其中，利用所述第一键矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二键矩阵采用动量梯度下降的方式进行更新包括：利用所述第一键矩阵和所述当前层 Transformer 网络在上一轮迭代得到的第二

键矩阵进行逐元素求差，得到键矩阵梯度；利用所述当前层 Transformer 网络在上一轮迭代得到的键矩阵动量和所述键矩阵梯度进行加权求和，得到所述当前层 Transformer 网络在当前轮迭代得到的键矩阵动量；利用所述当前层 Transformer 网络在当前轮迭代得到的键矩阵动量和当前层 Transformer 网络在上一轮迭代得到的第二键矩阵，得到所述更新后的第二键矩阵；

利用所述第一值矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二值矩阵采用动量梯度下降的方式进行更新包括：利用所述第一值矩阵和所述当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行逐元素求差，得到值矩阵梯度；利用所述当前层 Transformer 网络在上一轮迭代得到的值矩阵动量和所述值矩阵梯度进行加权求和，得到所述当前层 Transformer 网络在当前轮迭代得到的值矩阵动量；利用所述当前层 Transformer 网络在当前轮迭代得到的值矩阵动量和当前层 Transformer 网络在上一轮迭代得到的第二值矩阵，得到所述更新后的第二值矩阵。

5、根据权利要求 1 至 4 中任一项所述的方法，其中，每一轮迭代完成后，若确定达到预设的训练结束条件，则将各层 Transformer 网络当前迭代得到的第二键矩阵和第二值矩阵分别作为训练得到的各层 Transformer 网络的第二键矩阵和第二值矩阵进行存储。

6、一种文本预测方法，所述方法包括：

获取输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的输出标签；所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

7、根据权利要求 6 所述的方法，其中，所述大型语言模型还包括嵌入网络，所述文本预测模型还包括预测网络；

所述嵌入网络用以对所述文本序列进行嵌入处理；

若当前层 Transformer 网络为第一层 Transformer 网络，则所述上一层网络输出的特征表示为所述嵌入网络输出的特征表示；否则，所述上一层网络输出的特征表示为上一层 Transformer 网络输出的特征表示；

所述预测网络用以利用最后一层 Transformer 网络输出的特征表示预测输入文本对应的输出标签。

8、一种训练文本预测模型的方法，所述方法包括：

获取训练数据集，所述训练数据集包括输入文本样本以及所述输入文本样本对应的情感类别标签；

将包含输入文本样本和所述输入文本样本对应的情感类别标签的文本序列作为文本预测模型的输入，训练所述文本预测模型；其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的 Transformer 网络；

在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行：在当前轮迭代中，利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示；利用所述第一键矩阵对所述第二键矩阵进行更新，将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵；利用所述第一值矩阵对所述第二值矩阵进行更新，将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

9、根据权利要求 8 所述的方法，其中，所述大型语言模型还包括嵌入网络，所述文本预测模型还包括预测网络；

所述嵌入网络用以对所述文本序列进行嵌入处理；

若当前层 Transformer 网络为第一层 Transformer 网络，则所述上一层网络输出的特征表示为所述嵌入网络输出的特征表示；否则，所述上一层网络输出的特征表示为上一层 Transformer 网络输出的特征表示；

所述预测网络用以利用最后一层 Transformer 网络输出的特征表示预测输入文本样本对应的输出标签。

10、根据权利要求 8 所述的方法，其中，利用所述第一键矩阵对所述第二键矩阵进行更新包括：利用所述第一键矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二键矩阵采用动量梯度下降的方式进行更新；

利用所述第一值矩阵对所述第二值矩阵进行更新包括：利用所述第一值矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二值矩阵采用动量梯度下降的方式进行更新。

11、根据权利要求 10 所述的方法，其中，利用所述第一键矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二键矩阵采用动量梯度下降的方式进行更新包括：利用所述第一键矩阵和所述当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行逐元素求差，得到键矩阵梯度；利用所述当前层 Transformer 网络在上一轮迭代得到的键矩阵动量和所述键矩阵梯度进行加权求和，得到所述当前层

Transformer 网络在当前轮迭代得到的键矩阵动量；利用所述当前层 Transformer 网络在当前轮迭代得到的键矩阵动量和当前层 Transformer 网络在上一轮迭代得到的第二键矩阵，得到所述更新后的第二键矩阵；

利用所述第一值矩阵对所述当前层 Transformer 网络在上一轮迭代得到的第二值矩阵采用动量梯度下降的方式进行更新包括：利用所述第一值矩阵和所述当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行逐元素求差，得到值矩阵梯度；利用所述当前层 Transformer 网络在上一轮迭代得到的值矩阵动量和所述值矩阵梯度进行加权求和，得到所述当前层 Transformer 网络在当前轮迭代得到的值矩阵动量；利用所述当前层 Transformer 网络在当前轮迭代得到的值矩阵动量和当前层 Transformer 网络在上一轮迭代得到的第二值矩阵，得到所述更新后的第二值矩阵。

12、一种情感分析方法，所述方法包括：

获取输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的情感类别标签；所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

13、一种文本预测方法，由云端服务器执行，所述方法包括：

获取来自用户终端的输入文本；

将包含所述输入文本的文本序列输入文本预测模型，获取所述文本预测模型预测得到的所述输入文本对应的输出标签；

基于所述输出标签确定对应的服务内容，将所述服务内容发送至所述用户终端；

其中，所述文本预测模型采用大型语言模型，所述大型语言模型包括多层串连的转换 Transformer 网络；

各 Transformer 网络分别作为当前层 Transformer 网络执行：利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵；将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵，以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵；利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理，得到当前层 Transformer 网络输出的特征表示。

14、一种训练文本预测模型的装置，所述装置包括：

样本获取单元, 设置为获取训练数据集, 所述训练数据集包括输入文本样本以及所述输入文本样本对应的输出标签样本;

模型训练单元, 设置为将包含输入文本样本和所述输入文本样本对应的输出标签的文本序列作为文本预测模型的输入, 训练所述文本预测模型; 其中, 所述文本预测模型采用大型语言模型, 所述大型语言模型包括多层串连的转换 Transformer 网络; 在所述训练中各 Transformer 网络分别作为当前层 Transformer 网络执行: 在当前轮迭代中, 利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵; 将所述第一键矩阵与当前层 Transformer 网络在上一轮迭代得到的第二键矩阵进行拼接得到第三键矩阵, 以及将所述第一值矩阵与当前层 Transformer 网络在上一轮迭代得到的第二值矩阵进行拼接得到第三值矩阵; 利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理, 得到当前层 Transformer 网络输出的特征表示; 利用所述第一键矩阵对所述第二键矩阵进行更新, 将更新后的第二键矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二键矩阵; 利用所述第一值矩阵对所述第二值矩阵进行更新, 将更新后的第二值矩阵作为所述当前层 Transformer 网络在当前轮迭代得到的第二值矩阵。

15、一种文本预测装置, 所述装置包括:

文本获取单元, 设置为获取输入文本;

文本预测单元, 设置为将包含所述输入文本的文本序列输入文本预测模型, 获取所述文本预测模型预测得到的所述输入文本对应的输出标签; 所述文本预测模型采用大型语言模型, 所述大型语言模型包括多层串连的转换 Transformer 网络; 其中, 各 Transformer 网络分别作为当前层 Transformer 网络执行: 利用上一层网络输出的特征表示确定第一键矩阵、第一值矩阵和第一查询矩阵; 将所述第一键矩阵与预先训练得到的当前层 Transformer 网络的第二键矩阵进行拼接得到第三键矩阵, 以及将所述第一值矩阵与预先训练得到的当前层 Transformer 网络的第二值矩阵进行拼接得到第三值矩阵; 利用所述第三键矩阵、第三值矩阵和第一查询矩阵进行自注意力机制的处理, 得到当前层 Transformer 网络输出的特征表示。

16、一种计算机可读存储介质, 其上存储有计算机程序, 其中, 该程序被处理器执行时实现权利要求 1 至 13 中任一项所述的方法的步骤。

17、一种电子设备, 包括:

一个或多个处理器; 以及

与所述一个或多个处理器关联的存储器, 所述存储器设置为存储程序指令, 所述程序指令在被所述一个或多个处理器读取执行时, 执行权利要求 1 至 13 中任一项所述的方法的步骤。

18、一种计算机程序产品, 其中, 包括计算机程序, 所述计算机程序在被处理器执行时实现权利要求 1 至 13 中任意一项所述的方法的步骤。

19、一种计算机程序产品，其中，包括非易失性计算机可读存储介质，所述非易失性计算机可读存储介质存储计算机程序，所述计算机程序被处理器执行时实现权利要求 1 至 13 中任意一项所述的方法的步骤。

20、一种计算机程序，其中，所述计算机程序被处理器执行时实现权利要求 1 至 13 中任意一项所述的方法的步骤。