

多模态大语言模型训练方法、相关性计算及标签生成方法

5 本公开要求于 2023 年 12 月 14 日提交中国专利局、申请号为 202311724117.X、申请
名称为“多模态大语言模型训练方法、相关性计算及标签生成方法”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

10 本公开涉及人工智能技术领域，特别涉及一种多模态大语言模型训练方法、相关性计
算及标签生成方法。

背景技术

15 多模态大语言模型（Multimodal Large Language Mode，简称为 MLLM）是一种基于深
度学习的自然语言处理模型，能够对文本、图像、音频等多种类型数据进行处理，从而完
成诸如搜索、智能问答、翻译等自然语言任务。

20 目前，相关技术中的多模态大语言模型只能基于一种文本信息和图像信息生成多模态
特征向量。在搜索场景下，文本信息通常包括搜索词和商品的文本描述信息，而采用相关
技术所训练的多模态大语言模型只能学习到与搜索词有关的图像特征，生成与搜索词有关
的图像特征被增强的多模态特征向量，或者学习到与文本描述信息有关的图像特征，生成
与文本描述信息有关的图像特征被增强的多模态特征向量，无法同时学习到搜索场景中
与搜索词及文本描述信息有关的图像特征，导致所生成的多模态特征向量不够准确。因此，
亟需训练一种新的多模态大语言模型，从而生成更为准确的多模态特征向量，以满足搜索
场景下的应用需求。

25 发明内容

本公开实施例提供了一种多模态大语言模型训练方法、相关性计算及标签生成方法，
该方法所训练的多模态大语言模型能够同时学习到与搜索词和文本描述信息有关的图像
特征，生成的第二多模态特征向量更准确，满足了搜索场景的应用需求。所述技术方案如
下：

30 第一方面，提供了一种多模态大语言模型训练方法，所述方法包括：
获取样本搜索词和样本商品，所述样本商品具有图像信息和文本描述信息；
调用预训练多模态大语言模型，对所述样本商品的图像信息和文本描述信息进行处理，
得到所述样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，
所述样本第一多模态特征向量中与文本描述信息有关的图像特征被增强；

基于所述样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，对所述预训练多模态大语言模型进行训练，得到多模态大语言模型；

调用所述多模态大语言模型，对所述样本搜索词及所述样本商品的图像信息和文本描述信息进行处理，得到样本搜索词特征向量及所述样本商品的第一多模态特征向量和样本第二多模态特征向量，所述样本第二多模态特征向量中与文本描述信息和搜索词有关的图像特征被增强；

基于所述样本搜索词特征向量及所述样本商品的第一多模态特征向量和样本第二多模态特征向量，对所述多模态大语言模型进行训练，得到已训练的多模态大语言模型，所述已训练的多模态大语言模型用于基于搜索词及商品的图像信息和文本描述信息生成商品的第二多模态特征向量。

第二方面，提供了一种相关性计算方法，所述方法应用第一方面所述的已训练的多模态大语言模型，所述方法包括：

获取搜索词及基于所述搜索词搜索到的候选商品，所述候选商品具有图像信息和文本描述信息；

基于所述搜索词和所述文本描述信息，计算所述搜索词与所述候选商品之间的语义相关性分数；

调用所述已训练的多模态大语言模型，对所述搜索词、所述图像信息及文本描述信息进行处理，得到所述候选商品的第二多模态特征向量和搜索词特征向量；

基于所述搜索词特征向量和所述第二多模态特征向量，计算所述搜索词与所述候选商品之间的图文相关性分数；

基于所述语义相关性分数和所述图文相关性分数，计算所述搜索词与所述候选商品的总相关性分数。

在本公开的另一个实施例中，所述基于所述搜索词特征向量和所述第二多模态特征向量，计算所述搜索词与所述候选商品之间的图文相关性分数，包括：

计算所述搜索词特征向量和所述第二多模态特征向量之间的余弦相似度，得到所述图文相关性分数。

在本公开的另一个实施例中，所述基于所述语义相关性分数和所述图文相关性分数，计算所述搜索词与所述候选商品的总相关性分数，包括：

对所述语义相关性分数和所述图文相关性分数进行加权计算，得到所述总相关性分数。

第三方面，提供了一种标签生成方法，所述方法应用第一方面所述的已训练的多模态大语言模型，所述方法包括：

调用所述已训练的多模态大语言模型，对候选商品对应的搜索词、图像信息及文本描述信息进行处理，得到第二多模态特征向量；

获取所述候选商品所属的预设级别类目对应的第一指令模版，所述第一指令模版用于描述所述候选商品待输出的属性；

基于所述第二多模态特征向量和所述第一指令模版，生成所述候选商品在所述预设级别类目下的第一标签。

第四方面，提供了一种多模态大语言模型训练装置，所述装置包括：

5 第一获取模块，用于获取样本搜索词和样本商品，所述样本商品具有图像信息和文本描述信息；

第一处理模块，用于调用预训练多模态大语言模型，对所述样本商品的图像信息和文本描述信息进行处理，得到所述样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，所述样本第一多模态特征向量中与文本描述信息有关的图像特征被增强；

10 第一训练模块，用于基于所述样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，对所述预训练多模态大语言模型进行训练，得到多模态大语言模型；

第二处理模块，用于调用所述多模态大语言模型，对所述样本搜索词及所述样本商品的图像信息和文本描述信息进行处理，得到样本搜索词特征向量及所述样本商品的第一多模态特征向量和样本第二多模态特征向量，所述样本第二多模态特征向量中与文本描述信息和搜索词有关的图像特征被增强；

20 第二训练模块，用于基于所述样本搜索词特征向量及所述样本商品的第一多模态特征向量和样本第二多模态特征向量，对所述多模态大语言模型进行训练，得到已训练的多模态大语言模型，所述已训练的多模态大语言模型用于基于搜索词及商品的图像信息和文本描述信息生成商品的第二多模态特征向量。

第五方面，提供了一种相关性计算装置，所述装置应用第一方面所述的已训练的多模态大语言模型，所述装置包括：

获取模块，用于获取搜索词及基于所述搜索词搜索到的候选商品，所述候选商品具有图像信息和文本描述信息；

25 第一计算模块，用于基于所述搜索词和所述文本描述信息，计算所述搜索词与所述候选商品之间的语义相关性分数；

处理模块，用于调用所述已训练的多模态大语言模型，对所述搜索词、所述图像信息及文本描述信息进行处理，得到所述候选商品的第二多模态特征向量和搜索词特征向量；

30 第二计算模块，用于基于所述搜索词特征向量和所述第二多模态特征向量，计算所述搜索词与所述候选商品之间的图文相关性分数；

第三计算模块，用于基于所述语义相关性分数和所述图文相关性分数，计算所述搜索词与所述候选商品的总相关性分数。

第六方面，提供了一种标签生成装置，所述装置应用第一方面所述的已训练的多模态大语言模型，所述装置包括：

35 处理模块，用于调用所述已训练的多模态大语言模型，对候选商品对应的搜索词、图

像信息及文本描述信息进行处理，得到第二多模态特征向量；

第一获取模块，用于获取所述候选商品所属的预设级别类目对应的第一指令模版，所述
第一指令模版用于描述所述候选商品待输出的属性；

5 第一生成模块，用于基于所述第二多模态特征向量和所述第一指令模版，生成所述候
选商品在所述预设级别类目下的第一标签。

第七方面，提供了一种电子设备，包括处理器以及存储器；所述存储器存储至少一条
程序代码；所述至少一条程序代码用于被所述处理器调用并执行，以实现第一方面所述
的多模态大语言模型训练方法，或第二方面所述的相关性计算方法，或第三方面所述的标签
生成方法。

10 第八方面，提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有至少
一条计算机程序，所述至少一条计算机程序被处理器执行时能够实现第一方面所述的多模
态大语言模型训练方法，或第二方面所述的相关性计算方法，或第三方面所述的标签生成
方法。

第九方面，提供了一种计算机程序产品，所述计算机程序产品包括计算机程序，所述
15 计算机程序被处理器执行时能够实现第一方面所述的多模态大语言模型训练方法，或第二
方面所述的相关性计算方法，或第三方面所述的标签生成方法。

本公开实施例提供的技术方案带来的有益效果是：

采用两阶段的训练方式，对预训练多模态大语言模型进行训练，使得已训练的多模态
大语言模型同时学习到如何增强与搜索词和文本描述信息有关的图像特征。在第一训练阶
20 段，基于样本商品的图像信息和文本描述信息所生成的样本图像特征向量、样本文本特征
向量及样本第一多模态特征向量，对预训练多模态大语言模型进行训练，在训练过程中，
预训练多模态大语言模型学习到了如何增强与文本描述信息有关的图像特征。经过第一训
练阶段的训练，得到多模态大语言模型。在第二训练阶段，基于第一训练阶段训练得到的
多模态大语言模型，采用样本搜索词与样本商品的图像信息和文本描述信息，对该多模态
25 大语言模型进行训练，在训练过程中，该多模态大语言模型学习到了如何增强与搜索词和
文本描述信息有关的图像特征。经过第二训练阶段的训练，得到已训练的多模态大语言模
型，该已训练的多模态大语言模型所生成的第二多模态特征向量中与文本描述信息和搜索
词有关的图像特征均被增强，相比于只学习与文本描述信息有关的图像特征，或与搜索词
有关的图像特征，本公开所训练的多模态大语言模型学习到搜索场景中更全面的知识，生
30 成的第二模态特征向量更准确。

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，下面将对实施例或现有技
术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本公开
35 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 是本公开实施例提供的一种预训练多模态大语言模型的结构示意图；

图 2 是本公开实施例提供的一种基于多模态大语言模型在搜索场景的应用过程的示意图；

5 图 3 是本公开实施例提供的一种基于多模态大语言模型进行相关性计算的过程示意图；

图 4 是本公开实施例提供的一种基于多模态大语言模型生成商品标签的生成过程示意图；

图 5 是本公开实施例提供的多模态大语言模型训练方法、相关性计算方法及标签生成方法所涉及的实施环境的示意图；

10 图 6 是本公开实施例提供的一种多模态大语言模型训练方法的流程图；

图 7 是本公开实施例提供的一种相关性计算方法的流程图；

图 8 是本公开实施例提供的一种标签生成方法的流程图；

图 9 是本公开实施例提供的一种多模态大语言模型训练装置的结构示意图；

图 10 是本公开实施例提供的一种相关性计算装置的结构示意图；

15 图 11 是本公开实施例提供的一种标签生成装置的结构示意图；

图 12 示出了本公开一个示例性实施例提供的一种电子设备的结构框图。

具体实施方式

为使本公开的目的、技术方案和优点更加清楚，下面将结合附图对本公开实施方式作进一步地详细描述。

可以理解，本公开实施例所使用的术语“每个”、“多个”及“任一”等，多个包括两个或两个以上，每个是指对应的多个中的每一个，任一是指对应的多个中的任意一个。举例来说，多个词语包括 10 个词语，而每个词语是指这 10 个词语中的每一个词语，任一词语是指 10 个词语中的任意一个词语。

25 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。例如，本公开中涉及到的搜索词、用户偏好信息等都是在充分授权的情况下获取的。

在执行本公开实施例之前，首先对本公开实施例涉及的名词进行解释。

35 大语言模型（Large language Model，简称 LLM）是指使用大量文本数据训练的深度学习模型，可以生成自然语言文本或理解语言文本的含义。大语言模型可以处理多种自然语言任务，如文本分类、智能问答、对话等，是通向人工智能的一条重要途径。

Blip2 (Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, 用冻结的图像编码器和大型语言模型进行语言-图像预训练的自举模型) 是一种多模态 Transformer 模型, 主要用于解决 VLP (Vision-Language Pre-training, 视觉-语言预训练) 模型端到端训练导致计算代价过高的问题。

CLIP (Contrastive Language-Image Pre-Training, 对比语言图像预训练) 是一种基于对比的图片-文本学习的跨模态预训练模型。

指令/指示学习 (instruction learning) 用于对大语言模型进行提示或指令, 从而启发其获得更好的输出结果。

人工智能是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能, 感知环境、获取知识并使用知识获得结果的理论、方法、技术及应用系统。换句话说, 人工智能是计算机科学的一个综合技术, 它企图了解智能的实质, 并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法, 使机器具有感知、推理与决策的功能。

随着人工智能技术研究和进步, 人工智能技术在多个领域展开研究和应用, 例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服等, 相信随着技术的发展, 人工智能技术将在更多的领域得到应用, 并发挥越来越重要的价值。本公开实施例提供的方案涉及人工智能的自然语言处理等技术, 具体通过后续实施例进行说明。其中, 自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此, 这一领域的研究将涉及自然语言, 即人们日常使用的语言, 所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

在搜索场景下, 基于搜索词所搜索到的商品通常具有图像信息和文本描述信息 (例如, 标题、商品详情页信息、商品所属的类目、款式、型号等)。为便于后续应用, 会调用多模态大语言模型, 对商品的图像信息和相关的文本信息 (文本描述信息或搜索词) 进行融合处理, 以得到与文本信息有关的图像特征被增强的多模态特征向量。相关技术中的多模态大语言模型只进行一阶段的训练, 或者学习如何增强与搜索词有关的图像特征, 或者学习如何增强与文本描述信息有关的图像特征。实际上, 商品的图像不仅与搜索词有关, 还与商品本身的文本描述信息有关, 而相关技术中的多模态大语言模型只能生成与一种文本信息 (搜索词或文本描述信息) 有关的图像特征被增强的多模态融合特征向量, 因此, 相关技术所训练的多模态大语言模型不够准确, 无法满足搜索场景下的应用需求。

为此, 本公开实施例提供了一种多模态大语言模型的训练方法, 该方法采用两阶

段的训练方式，第一训练阶段，基于样本商品的图像信息和文本描述信息，对预训练多模态大语言模型进行训练，使得多模态大语言模型能够学习到如何增强与文本描述信息有关的图像特征，从而生成准确性较高的第一多模态特征向量。第二训练阶段，基于样本搜索词及样本商品的图像信息和文本描述信息，对第一训练阶段所训练的多模态特征向量进行训练，使得已训练的多模态大语言模型能够学习到如何增强与文本描述信息和搜索词有关的图像特征，具体来说，是学习如何对第一多模态特征向量中与搜索词有关的特征进行增强，从而生成准确性较高的第二多模态特征向量。

图 1 示出了本公开实施例所涉及的多模态大语言模型的一种结构示意图，参见图 1，该多模态大语言模型包括图像编码器、文本编码器及 Q-Former 结构等。

其中，图像编码器用于对图像信息进行编码，得到图像特征向量。图像信息包括商品的主图、主图的尺寸、对比度、色度等信息。文本编码器用于对文本描述信息（或搜索词）进行编码，得到文本特征向量（或搜索词特征向量）。图像编码器编码的图像特征向量与文本编码器编码的文本特征向量（或搜索词特征向量）的维度数相同，例如，均为 768 维等，以便于 Q-Former 结构对图像特征向量和文本特征向量（或搜索词特征向量）进行融合处理。

Q-Former 结构的输入为两种特征向量，输出为两种特征向量融合后的多模态特征向量，且输入的特征向量和输出的特征向量的维度数相同。Q-Former 结构包括至少一个自注意力模块、至少一个交叉注意力模块及至少一个前向反馈模块。自注意力模块用于计算输入的特征向量中每个元素与其他元素之间的相关性。交叉注意力模块用于在输入的两个特征向量之间计算注意力，以计算两个特征向量之间的关联关系。前向反馈模块用于增强 Q-Former 结构的非线性拟合能力。可以理解，图 1 中所示的 Q-Former 结构（包括两个交叉注意力模块、一个自注意力模块及两个前向反馈模块）只是一种示例性结构，可根据实际的计算需求，调整交叉注意力模块、自注意力模块及前向反馈模块的数量，并调整交叉注意力模块和自注意力模块的位置。

图 1 中所示的多模态大语言模型在不同的阶段的输入和输出是不同的，下面针对不同的阶段进行详细说明。

在第一训练阶段，对应的多模态大语言模型为预训练多模态大语言模型。对于任一样本商品，将该样本商品的图像信息输入到预训练多模态大语言模型的图像编码器中，经过该图像编码器进行编码，输出样本图像特征向量。将该样本商品的文本描述信息输入到预训练多模态大语言模型的文本编码器中，经过该文本编码器进行编码，输出样本文本特征向量。将样本图像特征向量和样本文本特征向量输入到预训练多模态大语言模型的 Q-Former 结构中，依次经过交叉注意力模块、自注意力模块、交叉注意力模块及前向反馈模块进行处理，输出样本第一多模态特征向量。然后，基于多个样本商品的样本第一多模态特征向量、样本图像特征向量及样本文本特征向量，对该预训练多模态大语言模型进行训练。

- 在第二训练阶段，对应的多模态大语言模型为第一训练阶段训练好的多模态大语言模型。对于任一样本商品，将该样本商品的图像信息输入到第一训练阶段训练好的多模态大语言模型的图像编码器中，经过该图像编码器进行编码，输出图像特征向量。将该样本图像的文本描述信息输入到第一训练阶段训练好的多模态大语言模型的文本编码器中，经过该文本编码器进行编码，输出文本特征向量。将图像特征向量和文本特征向量输入到第一训练阶段训练好的多模态大语言模型的 Q-Former 结构中，依次经过交叉注意力模块、自注意力模块、交叉注意力模块及前向反馈模块进行处理，输出第一多模态特征向量。接着，将样本搜索词输入到第一训练阶段训练好的多模态大语言模型的文本编码器中，经过该文本编码器进行编码，输出样本搜索词特征向量。接着，将样本搜索词特征向量和第一多模态特征向量输入到第一训练阶段训练好的多模态大语言模型的 Q-Former 结构中，依次经过交叉注意力模块、自注意力模块、交叉注意力模块及前向反馈模块进行处理，输出样本第二多模态特征向量。然后，基于多个样本商品的样本第二模态特征向量，对第一训练阶段训练好的多模态大语言模型进行训练，得到已训练的多模态大语言模型。
- 在搜索场景下，对于基于搜索词搜索的候选商品，可以计算搜索词与候选商品之间的相关性分数，从而按照相关性分数由高到低的顺序，对多个候选商品进行排序，以提升用户的点击率和转化率。目前，相关技术在计算搜索词与候选商品之间的相关性分数时，获取搜索词对应的搜索词特征向量，并获取候选商品的文本描述信息对应的文本特征向量，然后计算搜索词特征向量和文本特征向量之间的语义相关性分数，作为搜索词与候选商品之间的相关性分数。该方法只考虑了搜索词和候选商品的文本描述信息之间的语义相关性分数，相关性计算维度较为单一，计算结果不够准确。尤其是，存在候选商品的标题与图像不匹配，或者候选商品的标题是堆砌的、与候选商品的内容不符合，或者候选商品的标题过短等问题时，语义相关性分数的计算结果本身都不够可靠，更无法表征搜索词与候选商品之间的相关性。
- 为提高相关性计算结果的准确性，本公开实施例在原有的相关性计算的基础上，考虑了候选商品的图像信息和搜索词之间的图文相关性，增加了相关性计算维度。具体地，基于已训练的多模态大语言模型，生成候选商品的第二多模态特征向量和搜索词特征向量，进而基于第二多模态特征向量和搜索词特征向量，计算出候选商品和搜索词之间的图文相关性分数，然后对语义相关性分数和图文相关性分数进行权重计算，得到候选商品和搜索词之间的总相关性分数。本公开实施例的计算方法从多个角度表征候选商品与搜索词之间的相关性，相关性计算结果更可靠。更进一步地，还考虑了搜索词与用户偏好信息之间的正则相关性，然后通过对语义相关性分数、图文相关性分数及正则相关性分数进行加权计算，得到候选商品和搜索词之间的总相关性分数，相比于只是计算语义相关性分数，或者只是计算语义相关性分数和图文相关性分数，计算结果更为准确。

在搜索场景下，对于基于搜索词搜索的候选商品，还可以为搜索的候选商品生成合适的标签，所生成的标签可以为表示属性的词语，比如风格、款式、型号等，还可以是包含特定场景的短语，比如欧美风格的衣服、日韩风格的衣服、夏季的衣服、冬季的衣服等，从而更好地根据用户画像和搜索结果展示商品，以提升用户的点击率和转化率。相关技术在为候选商品生成标签时，通常基于深度神经网络训练一个判别式模型，然后通过调用该判别式模型，对候选商品的相关信息（例如，主图、标题、型号等）进行处理，得到每个候选标签的概率，进而将概率最大的标签作为该候选商品的标签。相关技术极大地依赖于对标签的人工预定义，而人工预定义的标签的数量及内容是固定的，无法根据搜索场景下用户的需求进行动态调整，可见，相关技术的标签生成方法不够灵活，具有一定的局限性。

为了能够根据用户的需求为候选商品灵活地生成标签，本公开实施例预先根据实际的应用场景，为不同级别类目设置不同的指令模版（prompt），并基于已训练的多模态大语言模型，生成候选商品的第二多模态特征向量，进而基于候选商品的某级类目对应的指令模版，对大语言模型进行指导，大语言模型根据指令模版进行指示学习（instruct learning），结合第二多模态特征向量所指示的候选商品的标题、图像等信息，为候选商品生成合适的标签。采用本公开实施例提供的方法无需为商品人工预定义标签，标签生成方式更灵活。

本公开实施例所训练的多模态大语言模型可用于计算基于搜索词搜索到的候选商品与搜索词之间的相关性，也可以用于为候选商品生成标签。图2示出了该已训练的多模态大语言模型的应用过程的示意图，参见图2，对于任一候选商品，将该候选商品的文本描述信息输入到已训练的多模态大语言模型的文本编码器中，输出该候选商品的文本特征向量，并将该候选商品的图像信息输入到已训练的多模态大语言模型的图像编码器中，输出该候选商品的图像特征向量。再将该候选商品的文本特征向量和图像特征向量进行融合，得到第一多模态特征向量。接着，将搜索词输入到已训练的多模态大语言模型的文本编码器中，输出搜索词特征向量。然后，将第一多模态特征向量和搜索词特征向量进行融合，得到第二多模态特征向量。基于所得到的第二多模态特征向量可进行相关性计算，也可以为候选商品生成标签。

需要说明的，上述以将文本特征向量和图像特征向量融合后，再将搜索词输入到文本编码器中为例进行说明的，也可以将文本描述信息和搜索词依次输入到文本编码器中，输出文本特征向量和搜索词特征向量之后，再将文本特征向量和图像特征向量进行融合，本公开实施例不对搜索词输入文本编码器的时机进行限定。

图3示出了本公开实施例提供的一种基于多模态大语言模型进行相关性计算的计算过程的示意图，参见图3，该相关性计算过程包括前置过程和应用过程。前置过程即为对预训练多模态大语言模型进行训练得到已训练的多模态大语言模型的过程，具体来说，采用样本搜索词及样本商品的图像信息和文本信息，对预训练大语言模型进

行训练，得到多模态大语言模型。应用过程即为基于已训练的多模态大语言模型所生成的搜索词特征向量和第二多模态特征向量，计算候选商品与搜索词的图文相关性分数，然后计算候选商品的总相关分数的过程，具体来说，当基于搜索词搜索到多个候选商品时，对于任一候选商品，可调用已训练的多模态大语言模型，对搜索词及该候选商品的图像信息和文本信息进行处理，得到搜索词特征向量和第二多模态特征向量，然后基于搜索词特征向量和第二多模态特征向量，计算候选商品与搜索词之间的图文相关性分数。与此同时，可基于搜索词和该候选商品的文本描述信息，计算候选商品与搜索词之间的语义相关性分数。还可以基于搜索词和用户偏好信息，计算出候选商品与搜索词之间的正则相关性分数。最后，对语义相关性分数、正则相关性分数及图文相关性分数进行加权计算，得到候选商品与搜索词之间的总相关性分数。

图4示出了本公开实施例提供的一种基于多模态大语言模型为候选商品生成标签的生成过程的示意图，参见图4，该标签生成过程包括前置过程和应用过程，前置过程即为对预训练多模态大语言模型进行训练得到已训练的多模态大语言模型的过程，具体来说，采用样本搜索词及样本商品的图像信息和文本信息，对预训练大语言模型进行训练，得到多模态大语言模型。应用过程即为基于指令模版和已训练的多模态大语言模型所生成的第二多模态特征向量，为候选商品生成标签的过程，具体来说，当基于搜索词搜索到多个候选商品时，对于任一候选商品，可调用已训练的多模态大语言模型，对搜索词及该候选商品的图像信息和文本信息进行处理，得到第二多模态特征向量，接着，获取该候选商品对应的指令模版，然后调用大语言模型，对第二多模态特征向量和指令模版进行处理，从而为该候选商品生成对应的标签。

请参考图5，其示出了本公开实施例提供的多模态大语言模型训练方法、相关性计算方法及标签生成方法所涉及的实施环境，参见图5，该实施环境包括：终端501、服务器502及服务器503。

其中，终端501可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等，但并不局限于此。服务器502和服务器503可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN（Content Delivery Network，内容分发网络）、以及大数据和人工智能平台等基础云计算服务的云服务器。上述终端501、服务器502及服务器503可以通过有线或无线通信方式进行直接或间接地连接，本公开实施例对此不作具体的限定。

对于多模态大语言模型训练过程来说，终端501用于获取用户输入的样本搜索词，向服务器502发送搜索请求，以便于服务器502返回搜索的多个样本商品，每个样本商品具有图像信息和文本信息。服务器502具有较强的计算能力，能够基于样本搜索词搜索多个样本商品。服务器503同样具有较强的计算能力，可基于样本搜索

词、多个样本商品的图像信息和文本信息，训练多模态大语言模型，对于具体的训练过程此处暂不详述。当多模态大语言模型训练完成后，可以直接部署到服务器 503 上，可以移植到其他服务器上。

对于相关性计算过程来说，终端 501 用于获取用户输入的搜索词，向服务器 502 发送搜索请求，以便于服务器 502 返回搜索的多个候选商品，每个候选商品具有图像信息和文本信息。服务器 502 具有较强的计算能力，能够基于搜索词搜索多个候选商品。如果服务器 502 上部署有包括已训练的多模态大语言模型在内的多个大语言模型，则服务器 502 可基于搜索词、每个候选商品的文本描述信息和图像信息、以及用户偏好信息，计算出搜索词与每个候选商品之间的总相关性分数，然后按照总相关性分数由高到低的顺序，将搜索到的多个候选商品返回给终端 501。如果服务器 502 上部署有除已训练的多模态大语言模型以外的多个大语言模型，而服务器 503 上部署有已训练的多模态大语言模型，则服务器 502 可调用服务器 503 上部署的多模态大语言模型，计算出搜索词与每个候选商品之间的图像相关性分数，并调用所部署的大语言模型，计算搜索词与每个候选商品之间的语义相关性分数，并基于用户偏好信息、搜索词等，计算出搜索词与每个候选商品之间的正则相关性分数，然后，基于搜索词与每个候选商品之间的语义相关性分数、图文相关性分数及图文相关性分数，计算搜索词与每个候选商品之间的总相关性分数。

对于标签生成过程来说，终端 501 用于获取用户输入的搜索词，向服务器 502 发送搜索请求，以便于服务器 502 返回搜索到的多个候选商品，每个候选商品具有图像信息和文本信息。服务器 502 具有较强的计算能力，能够基于搜索词搜索多个候选商品。如果服务器 502 上部署有包括已训练的多模态大语言模型在内的多个大语言模型，且存储有不同级别类目对应的指令模版，则服务器 502 可调用多模态大语言模型，对搜索词、每个候选商品的文本描述信息和图像信息进行处理，得到每个候选商品对应的第二多模态特征向量，并获取每个候选商品的某一级别类目对应的指令模版，然后，调用大语言模型，对每个候选商品对应的第二多模态特征向量和指令模版进行处理，得到每个候选商品的标签。如果服务器 502 上存储有不同级别类目对应的指令模版，但未部署已训练的多模态大语言模型，而服务器 503 上部署有已训练的多模态大语言模型，服务器 502 可调用服务器 503 上部署的多模态大语言模型，对搜索词、每个候选商品的文本描述信息和图像信息进行处理，得到每个候选商品对应的第二多模态特征向量，并获取每个候选商品的某一级别类目对应的指令模版，然后，调用大语言模型，对每个候选商品对应的第二多模态特征向量和指令模版进行处理，得到每个候选商品的标签。

基于图 5 所示的实施环境，本公开实施例提供了一种多模态大语言模型训练方法，以服务器 503 执行本公开实施例为例，参见图 6，本公开实施例提供的方法流程包括：

601、获取样本搜索词和样本商品。

其中，样本搜索词和样本商品为对多模态大语言模型进行训练的训练样本。样本搜索词可由用户输入，样本商品可基于样本搜索词搜索得到。在基于样本搜索词搜索样本商品时，可结合样本搜索词的原始文本、样本搜索词的语义向量、用户偏好信息、样本搜索词的改写词（即与样本搜索词语义相同或相近的词语）中至少一项进行搜索。由于本公开实施例所要训练的多模态大语言模型需要融合商品的图像信息和文本描述信息，因此，对于基于每个搜索词搜索到的商品，可采用人工方式或其他机器方式进行筛选，保留同时具有图像信息和文本描述信息的商品，删除具有图像信息而不具有文本描述信息，或者具有文本描述信息而不具有图像信息的商品，最终得到多个样本商品。

为便于对训练样本进行管理，可将每个样本搜索词对应的样本商品作为一个训练批次，并将多个样本搜索词对应的训练批次组成训练样本集合。每个训练批次包括样本搜索词对应的多个样本商品，这些样本商品包括正样本商品和负样本商品。不同训练阶段下，每个训练批次中包括的正样本商品和负样本商品的含义是不同。在第一训练阶段，正样本商品为图像信息和文本信息来源于同一商品的样本商品；负样本商品为图像信息和文本信息来源于不同商品的样本商品，该负样本商品可通过对多个样本搜索词搜索到的多个样本商品进行负采样得到，也即是，将一个样本搜索词搜索到的样本商品的图像信息与基于另一个样本搜索词搜索到的样本商品的文本描述信息组合得到。在第二训练阶段，正样本商品为与样本搜索词匹配的样本商品，即基于训练批次对应的样本搜索词搜索到的样本商品；负样本商品为与样本搜索词不匹配的样本商品，即不是基于训练批次对应的样本搜索词搜索到的样本商品，该负样本商品可通过对多个样本搜索词搜索到的多个样本商品进行负采样得到，也即是，将一个样本搜索词搜索到的样本商品与另一个样本搜索词组合得到。

进一步地，为了更好地完成本公开实施例对多模态大语言模型进行训练时所设置的训练任务（即目标函数），对于第一训练阶段和第二训练阶段，均可将一个正样本商品和多个负样本商品组成一个训练批次。

进一步地，为便于后续计算，对于每个样本商品，可对其对应的图像信息进行图像归一化处理，并对其对应的文本描述信息进行文本归一化处理。其中，图像归一化处理包括缩放、尺寸调整等。文本归一化处理包括删除停用词（例如，the、an、that等）、词形归一化（例如，将单复数统一）等。

602、调用预训练多模态大语言模型，对样本商品的图像信息和文本描述信息进行处理，得到样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量。

其中，预训练多模态大语言模型可以为 Blip2、CLIP 等。

具体地，调用预训练多模态大语言模型，对样本商品的图像信息和文本描述信息

进行处理，得到样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，包括：

6021、调用预训练多模态大语言模型，对样本商品的文本描述信息进行特征提取，得到样本商品的样本文本特征向量。

- 5 在本公开中预训练多模态大语言模型包括文本编码器，该文本编码器用于对文本描述信息进行编码。当将每个样本商品的文本描述信息输入到文本编码器中，文本编码器提取文本特征并进行编码，得到每个样本商品的样本文本特征向量。

6022、对样本商品的图像信息进行特征提取，得到样本商品的样本图像特征向量。

- 10 本公开的预训练多模态大语言模型包括图像编码器，该图像编码器用于对图像信息进行编码。当将每个样本商品的图像信息输入到图像编码器中，图像编码器提取图像特征并进行编码，得到每个样本商品的样本图像特征向量。

6023、将样本商品的样本文本特征向量和样本图像特征向量进行融合，得到样本商品的样本第一多模态特征向量。

- 15 本公开的预训练多模态大语言模型包括 Q-Former 结构，该 Q-Former 结构用于对输入的特征向量进行融合。对于每个样本商品，当将该样本商品的样本文本特征向量和样本图像特征向量输入到 Q-Former 结构中，样本文本特征向量和样本图像特征向量样本经过 Q-Former 结构中的交叉注意力模块、自注意力模块、前向反馈模块等进行处理，得到该样本商品的样本第一多模态特征向量，该样本第一多模态特征向量为
20 将样本文本特征向量和样本图像特征向量融合得到的特征向量，该样本第一多模态特征向量中与文本描述信息有关的图像特征被增强，当采用样本第一多模态特征向量对预训练多模态大语言模型进行训练，所训练的多模态大语言模型可以学习到如何对图像特征向量中与文本描述信息有关的图像特征进行增强。

- 25 603、基于样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，对预训练多模态大语言模型进行训练，得到多模态大语言模型。

其中，样本商品标注有表征图像信息与文本描述信息之间相关性的标签，包括相关标签和不相关标签。如果样本商品为正样本商品，则该样本商品标注的标签为相关标签；如果样本商品为负样本商品，则该样本商品标注的标签为不相关标签。

- 30 具体地，基于样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，对预训练多模态大语言模型进行训练，得到多模态大语言模型，包括：

6031、基于样本商品的样本第一多模态特征向量，获取表征样本商品的图像信息和文本描述信息之间相关性的标签。

- 35 本公开实施例在预训练多模态大语言模型的基础上，增加一个全连接层，该全连接层对应一个二分类训练任务，通过将每个样本商品的样本第一多模态特征向量输入到全连接层中，可输出表征每个样本商品的图像信息和文本描述信息之间相关性

的标签。

6032、基于第一总目标损失函数、样本商品的样本文本特征向量和样本图像特征向量、样本商品对应的生成标签和标注标签，对预训练多模态大语言模型进行训练，得到多模态大语言模型。

5 为更好地适配下游任务，本公开实施例针对两个训练阶段中的每个训练阶段均增加了两个训练任务，分别为图像文本匹配任务（Image-Text Matching，简称 ITM）和图像文本对比任务（Image-Text Contrastive Learning，简称 ITC），旨在建立图像和文本之间的匹配和对应关系，对齐图像和文本的表征，使其互信息最大化，从而得到多模态特征向量。

10 其中，图像文本匹配任务为一个二分类任务，要求模型能够预测出图像文本对匹配或不匹配。对于第一训练阶段来说，图像文本匹配任务的目标是每个样本商品对应的表征图像信息和文本描述信息之间相关性的标注标签与生成标签一致。对于第二训练阶段来说，图像文本匹配任务的目标是每个样本商品对应的表征样本搜索词和第一多模态特征向量相关性的标注标签与生成标签一致。

15 其中，图像文本对比任务的目标是让正样本的相似度尽量高，负样本的相似度尽量低。对于第一训练阶段来说，图像文本匹配任务的目标是让相关的样本图像特征向量和样本文本特征向量之间的距离尽可能接近，让不相关的样本图像特征向量和样本文本特征向量之间的距离尽可能远离；对于第二训练阶段来说，图像文本匹配任务的目标是让相关的样本搜索词特征向量和第一多模态特征向量之间的距离尽可能接近，
20 让不相关的样本搜索词特征向量和第一多模态特征向量之间的距离尽可能远离。

针对上述两个训练任务在两个训练阶段分别设置不同的总目标损失函数。其中，第一训练阶段对应第一总目标损失函数，第一总目标损失函数包括第一图像文本匹配损失函数和第一图像文本对比损失函数，该第一图像文本匹配损失函数需要完成第一训练阶段的图像文本匹配损失任务，该第一图像文本对比损失函数需要完成第一训练
25 阶段的图像文本对比学习任务；第二训练阶段对应第二总目标损失函数，第二总目标损失函数包括第二图像文本匹配损失函数和第二图像文本对比损失函数，该第二图像文本匹配损失函数需要完成第二训练阶段的图像文本匹配损失任务，该第二图像文本对比损失函数需要完成第二训练阶段的图像文本对比学习任务。

在第一训练阶段，基于第一总目标损失函数、样本商品的样本文本特征向量和样
30 本图像特征向量、样本商品对应的生成标签和标注标签，对预训练多模态大语言模型进行训练，得到多模态大语言模型，包括以下步骤：

60321、将样本商品对应的生成标签和标注标签输入到第一图像文本匹配损失函数中，得到第一图像文本匹配损失函数值。

60322、将样本文本特征向量和样本图像特征向量输入到第一图像文本对比损失
35 函数中，得到第一图像文本对比损失函数值。

60323、基于第一图像文本匹配损失函数值和第一图像文本对比损失函数值，确定第一总目标损失函数值。

获取第一图像文本匹配损失函数和第一图像文本对比损失函数在第一总目标损失函数中各自对应的权重值，对第一图像文本匹配损失函数值和第一图像文本对比损失函数值进行加权计算，得到第一总目标损失函数值。

60324、基于第一总目标损失函数值，对预训练多模态大语言模型的模型参数进行调整，得到多模态大语言模型。

第一训练阶段的优化目标是需要让第一总目标损失函数值最小，如果第一总目标损失函数值大于第一预设阈值，则需要调整预训练多模态大语言模型的模型参数，进而基于模型参数调整后的预训练多模态大语言模型，对每个样本商品的图像信息和文本信息进行处理，得到每个样本商品的样本图像特征向量、样本文本特征向量及样本第一多模态特征向量，然后基于每个样本商品的样本图像特征向量、样本文本特征向量及样本第一多模态特征向量，对模型参数调整后的预训练多模态大语言模型进行训练，直至得到的第一总目标损失函数的函数值小于第一预设阈值，或者，训练次数达到次数阈值。将训练截止时得到的预训练多模态大语言模型作为第一训练阶段训练好的多模态大语言模型（为便于后续描述，简称为多模态大语言模型），进而通过第二训练阶段对其进行再训练。

604、调用多模态大语言模型，对样本搜索词及样本商品的图像信息和文本描述信息进行处理，得到样本搜索词特征向量及每个样本商品的第一多模态特征向量和样本第二多模态特征向量。

其中，样本第二多模态特征向量中与文本描述信息和搜索词有关的图像特征被增强。在第二训练阶段中，以每个训练批次为单位，对第一阶段训练好的多模态大语言模型进行训练。每个训练批次包括多个样本商品，多个样本商品包括正样本商品和负样本商品。

具体地，调用多模态大语言模型，对样本搜索词及样本商品的图像信息和文本描述信息进行处理，得到样本搜索词特征向量及样本商品的第一多模态特征向量和样本第二多模态特征向量，包括：

6041、调用多模态大语言模型，对样本商品的文本描述信息进行特征提取，得到样本商品的文本特征向量。

第一训练阶段训练好的多模态大语言模型包括文本编码器，该文本编码器用于对文本描述信息进行编码。当将每个样本商品的文本描述信息输入到文本编码器中，文本编码器提取文本特征并进行编码，得到每个样本商品的文本特征向量。

6042、对样本商品的图像信息进行特征提取，得到样本商品的图像特征向量。

第一训练阶段训练好的多模态大语言模型包括图像编码器，该图像编码器用于对图像信息进行编码。当将每个样本商品的图像信息输入到图像编码器中，图像编码器

提取图像特征并进行编码，得到每个样本商品的图像特征向量。

6043、将样本商品的文本特征向量和图像特征向量进行融合，得到样本商品的第一多模态特征向量。

5 第一训练阶段训练好的多模态大语言模型包括 Q-Former 结构，该 Q-Former 结构用于对输入的特征向量进行融合。对于每个样本商品，当将该样本商品的文本特征向量和图像特征向量输入到 Q-Former 结构中，文本特征向量和图像特征向量样本经过 Q-Former 结构中的交叉注意力模块、自注意力模块、前向反馈模块等进行处理，得到该样本商品的第一多模态特征向量。

6044、对样本搜索词进行特征提取，得到样本搜索词特征向量。
10 将样本搜索词输入到第一训练阶段训练好的多模态大语言模型的文本编码器中，文本编码器提取搜索词特征并进行编码，得到样本搜索词特征向量。

6045、将样本搜索词特征向量与样本商品的第一多模态特征向量进行融合，得到样本商品的样本第二多模态特征向量。

15 将样本搜索词特征向量分别与每个样本商品的第一多模态特征向量输入到 Q-Former 结构中，样本搜索词特征向量和每个样本商品的第一多模态特征向量经过 Q-Former 结构中的交叉注意力模块、自注意力模块、前向反馈模块等进行处理，得到每个样本商品的样本第二多模态特征向量。

605、基于样本搜索词特征向量及样本商品的第一多模态特征向量和样本第二多模态特征向量，对多模态大语言模型进行训练，得到已训练的多模态大语言模型。

20 其中，每个样本商品标注有表征与样本搜索词之间相关性的标签。具体地，基于样本搜索词特征向量及每个样本商品的第一多模态特征向量和样本第二多模态特征向量，对多模态大语言模型进行训练，得到已训练的多模态大语言模型，包括以下步骤：

6051、基于每个样本商品的样本第二多模态特征向量，获取表征每个样本商品与
25 样本搜索词之间相关性的标签。

本公开实施例在多模态大语言模型的基础上，增加一个全连接层，该全连接层对应一个二分类训练任务，通过将每个样本商品的样本第二多模态特征向量输入到全连接层中，可输出表征每个样本商品与样本搜索词之间相关性的标签。

6052、基于第二总目标损失函数、每个样本商品的第一多模态特征向量和样本搜索词特征向量、每个样本商品对应的生成标签和标注标签，对多模态大语言模型进行
30 训练，得到已训练的多模态大语言模型。

其中，第二总目标损失函数包括第二图像文本匹配损失函数和第二图像文本对比损失函数。在第二训练阶段，基于第二总目标损失函数、每个样本商品的第一多模态特征向量和样本搜索词特征向量、每个样本商品对应的生成标签和标注标签，对多模
35 态大语言模型进行训练，得到已训练的多模态大语言模型，包括以下步骤：

60521、将每个样本商品对应的生成标签和标注标签输入到第二图像文本匹配损失函数中，得到第二图像文本匹配损失函数值。

60522、将每个样本商品的第一多模态特征向量和样本搜索词特征向量输入到第二图像文本对比损失函数中，得到第二图像文本对比损失函数值。

5 在本公开实施例中，第二图像文本对比损失函数可以表示为：

$$L_q = -\log \frac{\exp(q \cdot \frac{k_+}{t})}{\exp(q \cdot \frac{k_i}{t})}$$

其中， L_q 为第二图像文本对比损失函数， q 为搜索词特征向量， k_+ 为训练批次中正样本商品的第一多模态特征向量，即与搜索词特征向量相关的第一多模态特征向量， k_i 为训练批次中负样本商品的第一多模态特征向量，即与搜索词特征向量不相关
10 的第一多模态特征向量， t 为温度系数（自定义超参数）。

60523、基于第二图像文本匹配损失函数值和第二图像文本对比损失函数值，确定第二总目标损失函数值。

按照第二总目标损失函数中第二图像文本匹配损失函数和第二图像文本对比损失函数对应的权重值，对第二图像文本匹配损失函数值和第二图像文本对比损失函数
15 数值进行加权计算，得到第二总目标损失函数值。

60524、基于第二总目标损失函数值，对多模态大语言模型的模型参数进行调整，得到已训练的多模态大语言模型。

第二训练阶段的优化目标是让第二总目标损失函数值最小，如果第二总目标损失函数值大于第二预设阈值，则需要调整多模态大语言模型的模型参数，进而基于模型
20 参数调整后的多模态大语言模型，对样本搜索词及每个样本商品的图像信息和文本信息进行处理，得到样本搜索词特征向量、每个样本商品的第一多模态特征向量及样本第二多模态特征向量，然后基于样本搜索词特征向量、每个样本商品的第一多模态特征向量及样本第二多模态特征向量，对参数调整后的多模态大语言模型进行训练，直至得到的第二总目标损失函数的函数值小于第二预设阈值，或者，训练次数达到次数
25 阈值。将训练截止时得到的多模态大语言模型作为第二训练阶段训练好的多模态大语言模型，即本公开实施例所述的已训练的多模态大语言模型，该已训练的多模态大语言模型用于基于搜索词及商品的图像信息和文本描述信息生成商品的第二多模态特征向量。

经过上述两个训练阶段的学习，已训练的多模态大语言模型学习了如何从商品的
30 图像信息中提取与搜索词及文本描述信息有关的图像特征，调用该已训练的多模态大语言模型生成的第二多模态特征向量更准确，能够从多个角度反映商品图文融合后的特征。

上述所有可选技术方案，可以采用任意结合形成本公开的可选实施例，在此不再赘述。

基于图 5 所示的实施环境，本公开实施例提供了一种相关性计算方法，该方法应用上述图 6 中已训练的多模态大语言模型，以服务器 502 执行本公开实施例为例，参见图 7，本公开实施例提供的方法流程包括：

701、获取搜索词及基于搜索词搜索到的候选商品。

在电商搜索场景下，当用户想要购买某种商品时，用户可在用于提供该种商品购买服务的客户端（例如，应用程序、小程序、网页等）上输入用户自定义的、与该种商品相关的搜索词进行，响应于用户在客户端上的搜索操作，终端生成携带该搜索词的搜索请求，并将该搜索请求发送至服务器。响应于接收到用户发送的搜索词请求，服务器基于该搜索请求中携带的搜索词进行搜索，得到多个候选商品。每个候选商品具有图像信息和文本描述信息，可从不同角度对候选商品进行描述。在基于图像信息和文本描述信息对这些候选商品进行描述时发现，这些候选商品与搜索词具有不同的相关性，有的候选商品与搜索词的相关性要大一些，有的候选商品与搜索词的相关性要小一些，如果简单地按照搜索到的顺序展示给用户，可能会存在相关性较小的候选商品的展示位置比较靠前，相关性较大的候选商品的展示位置比较靠后，而用户则会按照展示位置由前到后浏览这些候选商品。由于展示位置靠前的候选商品与搜索词的相关性较小，无法满足用户的搜索需求，用户大概率不会继续浏览这些候选商品，导致候选商品的点击率和转化率较低。为了提高电商搜索领域的识别精度，可以计算搜索词与多个候选商品之间的相关性，从而按照相关性由高到低的顺序，将各个候选商品展示给用户，在提升用户体验的同时，为候选商品带来更多的用户点击率和转化率。

702、基于搜索词和文本描述信息，计算搜索词与候选商品之间的语义相关性分数。

其中，语义相关性分数用于反映搜索词与候选商品在语义角度上的相关程度。语义相关性分数越大，搜索词与候选商品在语义角度上的相关程度越高；语义相关性分数越小，搜索词与候选商品在语义角度上的相关程度越低。对于任一候选商品，在计算搜索词与该候选商品之间语义相关性分数时，可将搜索词和该候选商品的文本描述信息分别输入到大语言模型的文本编码器中，经过文本编码器的编码，输出搜索词特征向量和文本特征向量，进而计算搜索词特征向量和文本特征向量之间的相似度，然后将相似度计算结果作为搜索词与该候选商品之间的语义相关性分数。上述大语言模型可以为相关技术中任一具有文本编码器的大语言模型（此处不采用本公开实施例所训练的多模态大语言模型的原因在于，任何模型都有一定的计算精度，如果采用一个模型计算不同的相关性分数，则计算结果将受限于该模型的精度，如果采用不同的模型计算不同的相关性分数，则可通过调整不同模型对应的权重值，来提高相关性分数

的准确性)。在计算搜索词特征向量和文本特征向量之间的相似度时,可采用余弦相似度公式进行计算。

703、调用已训练的多模态大语言模型,对搜索词、图像信息及文本描述信息进行处理,得到候选商品的第二多模态特征向量和搜索词特征向量。

5 在本公开实施例中,第二多模态特征向量和搜索词特征向量的维度较高,为降低后续进行图文相关性计算时的计算量,可对第二多模态特征向量和搜索词特征向量进行降维处理。在一种可能的实现方式中,可在已训练的多模态大语言模型的基础上,增加一个全连接层,采用该全连接层对第二多模态特征向量和搜索词特征向量进行线性变换,能够直接输出维度较低的第二多模态特征向量和搜索词特征向量,例如,将
10 第二多模态特征向量和搜索词特征向量由 768 维降到 256 维。在另一种可能的实现方式中,可在对已训练的多模态大语言模型的结构进行改进,在上述已训练的多模态大语言模型中增加一个全连接层,当调用该已训练的多模态大语言模型对搜索词、图像信息及文本描述信息进行处理,直接得到维度较低的第二多模态特征向量和搜索词特征向量。当然,也可以采用其他方法对第二多模态特征向量和搜索词特征向量进行降
15 维处理,本公开实施例不再一一说明。

704、基于搜索词特征向量和第二多模态特征向量,计算搜索词与候选商品之间的图文相关性分数。

其中,图文相关性分数用于反映搜索词与候选商品在图文角度之间的相关程度,即搜索词的语义与候选商品的图像之间的相关程度。图文相关性分数越大,搜索词与
20 候选商品在图文角度上的相关程度越高;图文相关性分数越小,搜索词与候选商品在图文角度上的相关程度越低。在基于搜索词特征向量和第二多模态特征向量,计算搜索词与候选商品之间的图文相关性分数时,可以计算搜索词特征向量和第二多模态特征向量之间的余弦相似度,将得到的余弦相似度值作为图文相关性分数。其中,余弦相似度的计算公式为:

25

$$\text{Similarity} = \frac{AB}{|A||B|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

其中,Similarity为余弦相似度值, A 为搜索词特征向量, B 为第二多模态特征向量, A_i 为搜索词特征向量中第 i 维度上的元素, B_i 为第二多模态特征向量中第 i 维度上的元素, n 为搜索词特征向量(或第二多模态特征向量)的维度数。

30 705、基于语义相关性分数和图文相关性分数,计算搜索词与候选商品的总相关性分数。

在本公开实施例中,可以预先为语义相关性分数和图文相关性分数分别配置不同的权重值,进而基于语义相关性分数和图文相关性分数分别对应的权重值,对语义相

关性分数和图文相关性分数进行加权计算，得到搜索词与候选商品的总相关性分数。其中，为语义相关性分数和图文相关性分数配置的权重值并不是固定不变的，可根据在实际搜索场景中的应用结果进行动态调整。在为语义相关性分数和图文相关性分数配置不同的权重值时，包括但不限于如下两种方法：

5 第一种、可统计不同类目下语义相关性分数和图文相关性分数所占的比例，根据不同类目下语义相关性分数和图文相关性分数所在的比例进行离线拟合，得到不同类目下语义相关性分数和图文相关性分数分别对应的权重值。在应用时，可根据候选商品所属的类目，获取该类目下语义相关性分数和图文相关性分数分别对应的权重值。

10 第二种、可以训练一个用于确定语义相关性分数和图文相关性分数对应的权重值的模型。在应用时，可调用该模型确定语义相关性分数和图文相关性分数分别对应的权重值。

进一步地，考虑到不同的用户具有不同的偏好，用户的偏好也会影响到用户对候选商品的点击及转化，鉴于此，可以基于搜索词和用户偏好信息，计算搜索词与候选商品的正则相关性分数。正则相关性分数用于反映搜索词与候选商品在用户偏好角度上的相关程度。如果基于候选商品的文本描述信息确定出该候选商品符合用户的偏好，则可以确定搜索词与候选商品的正则相关性分数较高；如果基于候选商品的文本描述信息确定出该候选商品不符合用户的偏好，则可以确定搜索词与候选商品的正则相关性分数较低。基于所得到的语义相关性分数、图文相关性分数及正则相关性分数，通过对语义相关性分数、图文相关性分数及正则相关性分数进行加权计算，可得到搜索词与候选商品的总相关性分数。

20 相较于现有的一些电商场景下的相关性计算方法，本公开实施例基于多模态大语言模型，同时考虑了商品的图像信息、文本描述信息等，相关性打分的鲁棒性更高，计算结果更准确。由于相关性分数不再依赖于搜索词和文本描述信息之间的语义相关性分数，即便存在一些图文不匹配、商品标题堆砌或者过短问题，通过计算图文相关性分数及正则相关性分数，并调整不同相关性分数的权重值，也可以获得较好的计算结果。

上述所有可选技术方案，可以采用任意结合形成本公开的可选实施例，在此不再赘述。

30 基于图2所示的实施环境，本公开实施例提供了一种标签生成方法，该方法应用上述图6中已训练的多模态大语言模型，以图1中的服务器502执行本公开实施例为例，参见图8，本公开实施例提供的方法流程包括：

801、调用已训练的多模态大语言模型，对候选商品对应的搜索词、图像信息及文本描述信息进行处理，得到第二多模态特征向量。

35 在搜索场景下，对于搜索到的候选商品，通常会将这些候选商品添加标签，以便于后续可以根据用户画像和搜索结果进行展示，从而提高用户的点击率和转化率。对

于搜索到的候选商品，在为其生成标签之前，可对候选商品的图像信息进行缩放、尺寸变换等图像归一化处理，并对候选商品的文本描述信息进行删除停用词、大小写统一等文本归一化处理，从而作为后续模型的输入。

对于任一候选商品，调用已训练的多模态大语言模型，对该候选商品对应的搜索
5 词、图像信息及文本描述信息搜索词进行处理，可得到第二多模态特征向量。具体来说，可将文本描述信息中包括的标题、所属类目、商品详情页信息、型号等进行拼接，然后调用已训练的多模态大语言模型的文本编码器进行编码，得到文本特征向量，并调用已训练的多模态大语言模型的图像编码器对图像信息进行编码，得到图像特征向量，然后将文本特征向量和图像特征向量输入到已训练的多模态大语言模型的
10 Q-Former 中，输出第一多模态特征向量。接着，将搜索词输入到已训练的多模态大语言模型的文本编码器中进行编码，输出搜索词特征向量，然后将搜索词特征向量和第一多模态特征向量输入到已训练的多模态大语言模型的 Q-Former 中，输出该候选商品对应的第二多模态特征向量，该第二多模态特征向量中与搜索词和文本描述信息有关的图像特征被增强，能够反映出候选商品的各种属性。

进一步地，本公开实施例中，第二多模态特征向量的维度较高，为降低后续计算
15 时的计算量，可对第二多模态特征向量进行降维处理。在一种可能的实现方式中，可在已训练的多模态大语言模型的基础上，增加一个全连接层，采用该全连接层对第二多模态特征向量进行线性变换，能够直接输出维度较低的第二多模态特征向量，例如，将第二多模态特征向量由 768 维降到 256 维。在另一种可能的实现方式中，可在
20 对已训练的多模态大语言模型的结构进行改进，在上述已训练的多模态大语言模型中增加一个全连接层，当调用该已训练的多模态大语言模型对搜索词、图像信息及文本描述信息进行处理，直接得到维度较低的第二多模态特征向量。当然，也可以采用其他方法对第二多模态特征向量进行降维处理，本公开实施例不再一一说明。

需要说明的是，上述以为基于用户输入的搜索词搜索到候选商品生成标签为例，
25 如果需要生成标签的候选商品并不是基于用户输入的搜索词搜索到的，而是商家圈定的，则该候选商品对应的搜索词可以为服务器给出的、该候选商品最可能对应的搜索词。

802、获取候选商品所属的预设级别类目对应的第一指令模版。

考虑到不同类目下商品的标签是不同的，例如，服装类目的商品的标签可能是风格款式，运动类目的商品的标签可能是材质、重量、品牌等，针对不同的应用场景，
30 本公开实施例可对商品所属的不同级别类目进行划分，并为不同级别类目设置不同的指令模版，且每种级别类目对应的指令模版的数量为至少一个。其中，指令模版为经过设计的自然语言指令，用于描述候选商品待输出的属性，以指导模型生成相应属性。

35 例如，对于服装类目的商品，配置的指令模版可以为：

1. The material of this product is (请输出产品的材质), 该指令模版用于指导模型生成服装的材质;

2. Please give the style of the corresponding clothes according to the information (请根据提供的信息生成对应的产品风格), 该指令模版用于指导模型生成服装的风格。

又例如, 对于 3C 电子领域, 配置的指令模版可以为:

Please output the performance parameters of this product (请输出这个产品的性能参数), 该指令模版用于指导模型生成产品的性能参数。

进一步地, 为便于后续应用, 可存储不同级别类目与指令模版之间的对应关系。

在本公开实施例中, 对于任一候选商品, 可从该候选商品的文本描述信息中, 获取该候选商品所属的预设级别类目, 进而基于候选商品所属的预设级别类目, 从不同级别类目与指令模版之间的对应关系中, 获取该候选商品所属的预设级别类目对应的第一指令模版。其中, 预设级别类目根据需要设置, 可以为一级类目、二级类目等。

803、基于第二多模态特征向量和第一指令模版, 生成候选商品在预设级别类目下的第一标签。

在本公开实施例中, 第二多模态特征向量能够反映出候选商品的不同的属性, 第一指令模版用于指示需要待输出的候选商品的属性, 当将第二多模态特征向量和第一指令模版输入到大语言模型中, 大语言模型按照第一指令模版的指示, 通过多层 transformer 对第二多模态特征向量进行处理, 输出在第一指令模版指示的上下文中概率最大的词, 作为为候选商品生成的第一标签。该第一标签的数量与第一指令模版

的数量相同, 如果第一指令模版为一个, 则生成一个第一标签, 如果第一指令模版

的数量至少两个, 则生成至少两个标签。为便于将基于每个第一指令模版所生成的第一

标签区分开, 当基于每个第一指令模版输出在该第一指令模版所指示的上下文中概率

最大的词之后, 可在该词的词尾添加结束符, 以表示一次输出结束, 该结束符可以为

<EOF>。采用本公开实施例提供的方法, 最终可为每个候选商品生成至少一个第一标

签, 通过将这些第一标签进行组合, 可得到每个候选商品的最终标签。

考虑到大语言模型具有一定的精度问题, 为提高为候选商品生成的第一标签的可靠性, 可采用人工方式对候选商品对应的第一标签进行评估, 如果第一标签的评估结果为可靠, 则保留该候选商品对应的第一标签; 如果第一标签的评估结果为不可靠, 用户可重新为该候选商品所属的预设类目设置对应的指令模版, 即第二指令模版, 服务器获取用户针对预设级别类目输入的第二指令模版, 并将存储的对应关系中预设级别类目对应的第一指令模版更新为第二指令模版。可选地, 在评估过程中还可过滤掉重复标签, 避免重复标签对用户的干扰。

进一步地, 当将对应关系中预设级别类目对应的第一指令模版更新为第二指令模版之后, 还将基于候选商品的第二多模态特征向量和第二指令模版, 生成候选商品在

预设级别类目下的第二标签。具体地，可再次调用大语言模型，对候选商品的第二多模态特征向量和第二指令模版进行处理，得到候选商品在预设级别类目下的第二标签。当然，对于生成的第二标签也需要进行人工评估，如果评估结果为不可靠，需要用户再次为该候选商品所属的预设类目设置对应的指令模版，直至所生成的标签的评估结果为可靠。

本公开实施例可根据实际场景的需求，构建不同的指令模版。采用指令/指示学习方式，指导下游的大语言模型根据指令模版，结合商品的文本描述信息和图像信息等，为商品定向生成标签，从而极大节约了人工标注的成本。相较于传统的基于神经网络在已有模板库中进行匹配的方式，该方式生成的标签效果多样。

在搜索场景下，采用本公开实施例为商品生成的标签，同时结合用户画像，可将带有对应或者接近标签的产品展现给用户，从而提升用户的点击率和转化率。

上述所有可选技术方案，可以采用任意结合形成本公开的可选实施例，在此不再赘述。

请参考图 9，其示出了本公开实施例提供了一种多模态大语言模型训练装置的结构示意图，该装置可以通过软件、硬件或者二者结合实现，成为电子设备的全部或一部分，该装置包括：

第一获取模块 901，用于获取样本搜索词和样本商品，该样本商品具有图像信息和文本描述信息；

第一处理模块 902，用于调用预训练多模态大语言模型，对该样本商品的图像信息和文本描述信息进行处理，得到该样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，该样本第一多模态特征向量中与文本描述信息有关的图像特征被增强；

第一训练模块 903，用于基于该样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，对预训练多模态大语言模型进行训练，得到多模态大语言模型；

第二处理模块 904，用于调用多模态大语言模型，对样本搜索词及该样本商品的图像信息和文本描述信息进行处理，得到样本搜索词特征向量及该样本商品的第一多模态特征向量和样本第二多模态特征向量，该样本第二多模态特征向量中与文本描述信息和搜索词有关的图像特征被增强；

第二训练模块 905，用于基于样本搜索词特征向量及该样本商品的第一多模态特征向量和样本第二多模态特征向量，对多模态大语言模型进行训练，得到已训练的多模态大语言模型，该已训练的多模态大语言模型用于基于搜索词及商品的图像信息和文本描述信息生成商品的第二多模态特征向量。

在本公开的另一个实施例中，第一处理模块 902，用于调用预训练多模态大语言模型，对该样本商品的文本描述信息进行特征提取，得到该样本商品的样本文本特征

向量；对该样本商品的图像信息进行特征提取，得到该样本商品的样本图像特征向量；将该样本商品的样本文本特征向量和样本图像特征向量进行融合，得到该样本商品的样本第一多模态特征向量。

5 在本公开的另一个实施例中，该样本商品标注有表征图像信息与文本描述信息之间相关性的标签，第一训练模块 903，用于基于该样本商品的样本第一多模态特征向量，获取表征该样本商品的图像信息和文本描述信息之间相关性的标签；基于第一总目标损失函数、该样本商品的样本文本特征向量和样本图像特征向量、该样本商品对应的生成标签和标注标签，对预训练多模态大语言模型进行训练，得到多模态大语言模型。

10 在本公开的另一个实施例中，第一总目标损失函数包括第一图像文本匹配损失函数和第一图像文本对比损失函数，第一训练模块 903，用于将该样本商品对应的生成标签和标注标签输入到第一图像文本匹配损失函数中，得到第一图像文本匹配损失函数值；将该样本文本特征向量和样本图像特征向量输入到第一图像文本对比损失函数中，得到第一图像文本对比损失函数值；基于第一图像文本匹配损失函数值和第一图
15 像文本对比损失函数值，确定第一总目标损失函数值；基于第一总目标损失函数值，对预训练多模态大语言模型的模型参数进行调整，得到多模态大语言模型。

在本公开的另一个实施例中，第二处理模块 904，用于调用多模态大语言模型，对该样本商品的文本描述信息进行特征提取，得到该样本商品的文本特征向量；对该样本商品的图像信息进行特征提取，得到该样本商品的图像特征向量；将该样本商品
20 的文本特征向量和图像特征向量进行融合，得到该样本商品的样本第一多模态特征向量；对样本搜索词进行特征提取，得到样本搜索词特征向量；将样本搜索词特征向量分别与该样本商品的样本第一多模态特征向量进行融合，得到该样本商品的样本第二多模态特征向量。

在本公开的另一个实施例中，该样本商品标注有表征与样本搜索词之间相关性的
25 标签，第二训练模块 905，用于基于该样本商品的样本第二多模态特征向量，获取表征该样本商品与样本搜索词之间相关性的标签；基于第二总目标损失函数、该样本商品的样本第一多模态特征向量和样本搜索词特征向量、该样本商品对应的生成标签和标注标签，对多模态大语言模型进行训练，得到已训练的多模态大语言模型。

在本公开的另一个实施例中，第二总目标损失函数包括第二图像文本匹配损失函数和第二图像文本对比损失函数，第二训练模块 905，用于将该样本商品对应的生成
30 标签和标注标签输入到第二图像文本匹配损失函数中，得到第二图像文本匹配损失函数值；将该样本商品的样本第一多模态特征向量和样本搜索词特征向量输入到第二图像文本对比损失函数中，得到第二图像文本对比损失函数值；基于第二图像文本匹配损失函数值和第二图像文本对比损失函数值，确定第二总目标损失函数值；基于第二总目
35 标损失函数值，对多模态大语言模型的模型参数进行调整，得到已训练的多模态大语

言模型。

请参考图 10，其示出了本公开实施例提供了一种相关性计算装置的结构示意图，该装置应用图 6 中已训练的多模态大语言模型，该装置可以通过软件、硬件或者二者结合实现，成为电子设备的全部或一部分，该装置包括：

5 获取模块 1001，用于获取搜索词及基于搜索词搜索到的候选商品，候选商品具有图像信息和文本描述信息；

第一计算模块 1002，用于基于搜索词和文本描述信息，计算搜索词与候选商品之间的语义相关性分数；

10 处理模块 1003，用于调用已训练的多模态大语言模型，对搜索词、图像信息及文本描述信息进行处理，得到候选商品的第二多模态特征向量和搜索词特征向量；

第二计算模块 1004，用于基于搜索词特征向量和第二多模态特征向量，计算搜索词与候选商品之间的图文相关性分数；

第三计算模块 1005，用于基于语义相关性分数和图文相关性分数，计算搜索词与候选商品的总相关性分数。

15 在本公开的另一个实施例中，第二计算模块 1004，用于计算搜索词特征向量和第二多模态特征向量之间的余弦相似度，得到图文相关性分数。

在本公开的另一个实施例中，第三计算模块 1005，用于对语义相关性分数和图文相关性分数进行加权计算，得到总相关性分数。

在本公开的另一个实施例中，该装置还包括：

20 第四计算模块，用于基于文本描述信息与用户偏好信息，计算搜索词与商品的正则相关性分数；

第五计算模块，用于对语义相关性分数、图文相关性分数及正则相关性分数进行加权计算，得到搜索词与候选商品的总相关性分数。

25 请参考图 11，其示出了本公开实施例提供了一种标签生成装置的结构示意图，该装置应用图 6 中已训练的多模态大语言模型，该装置可以通过软件、硬件或者二者结合实现，成为电子设备的全部或一部分，该装置包括：

处理模块 1101，用于调用已训练的多模态大语言模型，对候选商品对应的搜索词、图像信息及文本描述信息进行处理，得到第二多模态特征向量；

30 第一获取模块 1102，用于获取候选商品所属的预设级别类目对应的第一指令模版，该第一指令模版用于描述候选商品待输出的属性；

第一生成模块 1103，用于基于第二多模态特征向量和第一指令模版，生成候选商品在预设级别类目下的第一标签。

在本公开的另一个实施例中，第一获取模块 1102，用于基于候选商品所属的预设级别类目，从不同级别类目与指令模版之间的对应关系中，获取第一指令模版。

35 在本公开的另一个实施例中，该装置还包括：

第二获取模块，用于如果第一标签的评估结果为不可靠，获取用户针对预设级别类目输入的第二指令模版；

更新模块，用于将对应关系中预设级别类目对应的第一指令模版更新为第二指令模版。

5 在本公开的另一个实施例中，该装置还包括：

第二生成模块，用于基于第二多模态特征向量和第二指令模版，生成候选商品在预设级别类目下的第二标签。

图 12 示出了本公开一个示例性实施例提供的一种电子设备 1200 的结构框图。通常，电子设备 1200 包括有：处理器 1201 和存储器 1202。

10 处理器 1201 可以采用 DSP（Digital Signal Processing，数字信号处理）、FPGA（Field-Programmable Gate Array，现场可编程门阵列）、PLA（Programmable Logic Array，可编程逻辑阵列）中的至少一种硬件形式来实现。处理器 1201 也可以包括主处理器和协处理器，主处理器是用于对在唤醒状态下的数据进行处理的处理

15 器 1201 可以在集成有 GPU（Graphics Processing Unit，图像处理器），GPU 用于负责显示屏所需要显示的内容的渲染和绘制。一些实施例中，处理器 1201 还可以包括人工智能处理器，该人工智能处理器用于处理有关机器学习的计算操作。

存储器 1202 可以包括一个或多个计算机可读存储介质，该计算机可读存储介质可以是是非临时性计算机可读存储介质，例如，所述非临时性计算机可读存储介质可以是 CD-ROM（Compact Disc Read-Only Memory，只读光盘）、ROM、RAM（Random Access Memory，随机存取存储器）、磁带、软盘和光数据存储设备等。

20 该计算机可读存储介质中存储有至少一条计算机程序，该至少一条计算机程序被执行时能够实现上述多模态大语言模型生成方法，或相关性计算方法，或标签生成方法。

当然，上述电子设备必然还可以包括其他部件，例如输入/输出接口、通信组件等。输入/输出接口为处理器和外围接口模块之间提供接口，上述外围接口模块可以是输出设备、输入设备等。通信组件被配置为便于电子设备和

25 其他设备之间有线或无线方式的通信等。

本领域技术人员可以理解，图 12 中示出的结构并不构成对电子设备 1200 的限定，可以包括比图示更多或更少的组件，或者组合某些组件，或者采用不同的组件布置。

30 本公开实施例提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一条计算机程序，所述至少一条计算机程序被处理器执行时能够实现上述多模态大语言模型训练方法，或相关性计算方法，或标签生成方法。

本公开实施例提供了一种计算机程序产品，所述计算机程序产品包括计算机程序，所述计算机程序被处理器执行时能够实现上述多模态大语言模型训练方法，或相关性计算方法，或标签生成方法。

35

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

- 5 以上实施例仅用以说明本公开的技术方案，而非对其限制；尽管参照前述实施例对本公开进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本公开各实施例技术方案的精神和范围。

权 利 要 求 书

- 1、一种多模态大语言模型训练方法，其中，所述方法包括：
- 获取样本搜索词和样本商品，所述样本商品具有图像信息和文本描述信息；
- 调用预训练多模态大语言模型，对所述样本商品的图像信息和文本描述信息进行处理，得到所述样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，所述样本第一多模态特征向量中与文本描述信息有关的图像特征被增强；
- 基于所述样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，对所述预训练多模态大语言模型进行训练，得到多模态大语言模型；
- 调用所述多模态大语言模型，对所述样本搜索词及所述样本商品的图像信息和文本描述信息进行处理，得到样本搜索词特征向量及所述样本商品的第一多模态特征向量和样本第二多模态特征向量，所述样本第二多模态特征向量中与文本描述信息和搜索词有关的图像特征被增强；
- 基于所述样本搜索词特征向量及所述样本商品的第一多模态特征向量和样本第二多模态特征向量，对所述多模态大语言模型进行训练，得到已训练的多模态大语言模型，所述已训练的多模态大语言模型用于基于搜索词及商品的图像信息和文本描述信息生成商品的第二多模态特征向量。
- 2、根据权利要求1所述的方法，其中，所述调用预训练多模态大语言模型，对所述样本商品的图像信息和文本描述信息进行处理，得到所述样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，包括：
- 调用所述预训练多模态大语言模型，对所述样本商品的文本描述信息进行特征提取，得到所述样本商品的样本文本特征向量；
- 对所述样本商品的图像信息进行特征提取，得到所述样本商品的样本图像特征向量；
- 将所述样本商品的样本文本特征向量和样本图像特征向量进行融合，得到所述样本商品的样本第一多模态特征向量。
- 3、根据权利要求1所述的方法，其中，所述样本商品标注有表征图像信息与文本描述信息之间相关性的标签，所述基于所述样本商品的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，对所述预训练多模态大语言模型进行训练，得到多模态大语言模型，包括：
- 基于所述样本商品的样本第一多模态特征向量，获取表征所述样本商品的图像信息和文本描述信息之间相关性的标签；
- 基于第一总目标损失函数、所述样本商品的样本文本特征向量和样本图像特征向

量、所述样本商品对应的生成标签和标注标签，对所述预训练多模态大语言模型进行训练，得到所述多模态大语言模型。

- 4、根据权利要求3所述的方法，其中，所述第一总目标损失函数包括第一图像文本匹配损失函数和第一图像文本对比损失函数，所述基于所述第一总目标损失函数、所述样本商品的样本文本特征向量和样本图像特征向量、所述样本商品对应的生成标签和标注标签，对所述预训练多模态大语言模型进行训练，得到所述多模态大语言模型，包括：

10 将所述样本商品对应的生成标签和标注标签输入到所述第一图像文本匹配损失函数中，得到第一图像文本匹配损失函数值；

将所述样本文本特征向量和样本图像特征向量输入到所述第一图像文本对比损失函数中，得到第一图像文本对比损失函数值；

基于所述第一图像文本匹配损失函数值和所述第一图像文本对比损失函数值，确定第一总目标损失函数值；

15 基于所述第一总目标损失函数值，对所述预训练多模态大语言模型的模型参数进行调整，得到所述多模态大语言模型。

- 5、根据权利要求1所述的方法，其中，所述调用所述多模态大语言模型，对所述样本搜索词及所述样本商品的图像信息和文本描述信息进行处理，得到样本搜索词特征向量及所述样本商品的第一多模态特征向量和样本第二多模态特征向量，包括：

调用所述多模态大语言模型，对所述样本商品的文本描述信息进行特征提取，得到所述样本商品的文本特征向量；

对所述样本商品的图像信息进行特征提取，得到所述样本商品的图像特征向量；

25 将所述样本商品的文本特征向量和图像特征向量进行融合，得到所述样本商品的第一多模态特征向量；

对所述样本搜索词进行特征提取，得到所述样本搜索词特征向量；

将所述样本搜索词特征向量与所述样本商品的第一多模态特征向量进行融合，得到所述样本商品的样本第二多模态特征向量。

- 30 6、根据权利要求1所述的方法，其中，所述样本商品标注有表征与所述样本搜索词之间相关性的标签，所述基于所述样本搜索词特征向量及所述样本商品的第一多模态特征向量和样本第二多模态特征向量，对所述多模态大语言模型进行训练，得到已训练的多模态大语言模型，包括：

35 基于所述样本商品的样本第二多模态特征向量，获取表征所述样本商品与所述样本搜索词之间相关性的标签；

基于第二总目标损失函数、所述样本商品的第一多模态特征向量和样本搜索词特征向量、所述样本商品对应的生成标签和标注标签，对所述多模态大语言模型进行训练，得到已训练的多模态大语言模型。

5 7、根据权利要求6所述的方法，其中，所述第二总目标损失函数包括第二图像文本匹配损失函数和第二图像文本对比损失函数，所述基于第二总目标损失函数、所述样本商品的第一多模态特征向量和样本搜索词特征向量、所述样本商品对应的生成标签和标注标签，对所述多模态大语言模型进行训练，得到已训练的多模态大语言模型，包括：

10 将所述样本商品对应的生成标签和标注标签输入到所述第二图像文本匹配损失函数中，得到第二图像文本匹配损失函数值；

 将所述样本商品的第一多模态特征向量和样本搜索词特征向量输入到所述第二图像文本对比损失函数中，得到第二图像文本对比损失函数值；

15 基于所述第二图像文本匹配损失函数值和所述第二图像文本对比损失函数值，确定第二总目标损失函数值；

 基于所述第二总目标损失函数值，对所述多模态大语言模型的模型参数进行调整，得到已训练的多模态大语言模型。

20 8、一种相关性计算方法，其中，所述方法应用权利要求1至7中任一项所述的已训练的多模态大语言模型，所述方法包括：

 获取搜索词及基于所述搜索词搜索到的候选商品，所述候选商品具有图像信息和文本描述信息；

 基于所述搜索词和所述文本描述信息，计算所述搜索词与所述候选商品之间的语义相关性分数；

25 调用所述已训练的多模态大语言模型，对所述搜索词、所述图像信息及文本描述信息进行处理，得到所述候选商品的第二多模态特征向量和搜索词特征向量；

 基于所述搜索词特征向量和所述第二多模态特征向量，计算所述搜索词与所述候选商品之间的图文相关性分数；

30 基于所述语义相关性分数和所述图文相关性分数，计算所述搜索词与所述候选商品的总相关性分数。

 9、根据权利要求8所述的方法，其中，所述方法还包括：

 基于所述文本描述信息与用户偏好信息，计算所述搜索词与所述候选商品的正则相关性分数；

35 对所述语义相关性分数、所述图文相关性分数及所述正则相关性分数进行加权计算，得到所述总相关性分数。

10、一种标签生成方法，其中，所述方法应用权利要求 1 至 7 中任一项所述的已训练的多模态大语言模型，所述方法包括：

5 调用所述已训练的多模态大语言模型，对候选商品对应的搜索词、图像信息及文本描述信息进行处理，得到第二多模态特征向量；

获取所述候选商品所属的预设级别类目对应的第一指令模版，所述第一指令模版用于描述所述候选商品待输出的属性；

基于所述第二多模态特征向量和所述第一指令模版，生成所述候选商品在所述预设级别类目下的第一标签。

10

11、根据权利要求 10 所述的方法，其中，所述获取所述候选商品所属的预设级别类目对应的第一指令模版，包括：

基于所述候选商品所属的预设级别类目，从不同级别类目与指令模版之间的对应关系中，获取所述第一指令模版。

15

12、根据权利要求 11 所述的方法，其中，所述基于所述第二多模态特征向量和所述第一指令模版，生成所述候选商品在所述预设级别类目下的第一标签之后，还包括：

20 如果所述第一标签的评估结果为不可靠，获取用户针对所述预设级别类目输入的第二指令模版；

将所述对应关系中所述预设级别类目对应的所述第一指令模版更新为所述第二指令模版。

25 13、根据权利要求 12 所述的方法，其中，所述将所述对应关系中所述预设级别类目对应的所述第一指令模版更新为所述第二指令模版之后，还包括：

基于所述第二多模态特征向量和所述第二指令模版，生成所述候选商品在所述预设级别类目下的第二标签。

30 14、一种电子设备，其中，包括处理器以及存储器；所述存储器存储至少一条程序代码；所述至少一条程序代码用于被所述处理器调用并执行，以实现如权利要求 1 至 7 中任一项所述的多模态大语言模型训练方法，或权利要求 8 或 9 所述的相关性计算方法，或权利要求 10 至 13 中任一项所述的标签生成方法。

35 15、一种计算机可读存储介质，其中，所述计算机可读存储介质中存储有至少一条计算机程序，所述至少一条计算机程序被处理器执行时能够实现如权利要求 1 至 7

中任一项所述的多模态大语言模型训练方法，或权利要求 8 或 9 所述的相关性计算方法，或权利要求 10 至 13 中任一项所述的标签生成方法。

- 5 16、一种计算机程序产品，其中，所述计算机程序产品包括计算机程序，所述计算机程序被处理器执行时能够实现如权利要求 1 至 7 中任一项所述的多模态大语言模型训练方法，或权利要求 8 或 9 所述的相关性计算方法，或权利要求 10 至 13 中任一项所述的标签生成方法。