

多模态预训练方法和装置

相关申请的交叉引用

本公开是以 CN 申请号为 202111078728.2，申请日为 2021 年 9 月 15 日的申请为
5 基础，并主张其优先权，该 CN 申请的公开内容在此作为整体引入本公开中。

技术领域

本公开涉及信息处理领域，特别涉及一种多模态预训练方法和装置。

10 背景技术

视觉语言多模态预训练技术是近期多模态领域的新兴课题之一，它的目的是让模型能对进行大规模弱标注的视觉（如图像、视频）和文本数据进行预训练以得到一个更好的多模态特征表示，从而提升各种多模态任务模型的性能。

视觉语言多模态预训练的相关技术基本都是借鉴自然语言处理领域的 **BERT**
15 （**Bidirectional Encoder Representations From Transformer**，基于变换器的双向编码器表征）预训练模型的方法。

发明内容

根据本公开实施例的第一方面，提供一种多模态预训练方法，包括：对视频-文本
20 对中的视频进行采样，以得到第一视频帧序列；对所述视频-文本对中的文本进行分词处理，以得到第一分词序列；对所述第一视频帧序列进行掩码处理，以得到第二视频帧序列；对所述第一分词序列进行掩码处理，以得到第二分词序列；对所述第一视频帧序列进行编码，以得到第一视频特征，对所述第一分词序列进行编码，以得到第一分词特征；对所述第二视频帧序列进行编码，以得到第二视频特征，对所述第二分词序列进行编码，以得到第二分词特征；利用所述第一视频特征、所述第一分词特征、
25 所述第二视频特征、所述第二分词特征确定预训练的目标函数；利用所述预训练的目标函数进行多模态预训练。

在一些实施例中，确定预训练的目标函数包括：利用所述第一分词特征、所述第二视频特征和预设的第一负样本特征确定第一对比损失值；利用所述第一视频特征、
30 所述第二分词特征和预设的第二负样本特征确定第二对比损失值；根据所述第一对比

损失值和所述第二对比损失值确定第一目标；利用所述第一视频特征、所述第二视频特征和所述第二负样本特征确定第三对比损失值；利用所述第一分词特征、所述第二分词特征和所述第一负样本特征确定第四对比损失值；根据所述第三对比损失值和所述第四对比损失值确定第二目标；根据所述第一目标和所述第二目标确定所述目标函数。

在一些实施例中，确定第一对比损失值包括：将所述第一分词特征转换为全局的第一正样本特征；将所述第二视频特征转换为全局的视频查询特征；利用所述视频查询特征、所述第一正样本特征和所述第一负样本特征确定第一对比损失值。

在一些实施例中，确定第二对比损失值包括：将所述第一视频特征转换为全局的第二正样本特征；将所述第二分词特征转换为全局的文本查询特征；利用所述文本查询特征、所述第二正样本特征和所述第二负样本特征确定第二对比损失值。

在一些实施例中，确定第三对比损失值包括：利用所述视频查询特征、所述第二正样本特征和所述第二负样本特征确定第三对比损失值。

在一些实施例中，确定第四对比损失值包括：利用所述文本查询特征、所述第一正样本特征和所述第一负样本特征确定第四对比损失值。

在一些实施例中，所述第一目标为所述第一对比损失值和所述第二对比损失值之和；所述第二目标为所述第三对比损失值和所述第四对比损失值之和。

在一些实施例中，所述目标函数为所述第一目标和所述第二目标之和。

在一些实施例中，对所述第二视频特征和所述第二分词特征进行融合处理，以得到融合特征；将所述融合特征输入带掩码的文本建模 MLM 模型以得到第三目标，将所述融合特征输入带掩码的文本生成 MSG 模型以得到第四目标；所述根据所述第一目标和所述第二目标确定所述目标函数包括：根据所述第一目标、所述第二目标、所述第三目标和所述第四目标确定所述目标函数。

在一些实施例中，所述目标函数为所述第一目标、所述第二目标、所述第三目标和所述第四目标之和。

根据本公开实施例的第二方面，提供一种多模态预训练装置，包括：第一处理模块，被配置为对视频-文本对中的视频进行采样，以得到第一视频帧序列，还被配置为对所述视频-文本对中的文本进行分词处理，以得到第一分词序列；第二处理模块，被配置为对所述第一视频帧序列进行掩码处理，以得到第二视频帧序列，还被配置为对所述第一分词序列进行掩码处理，以得到第二分词序列；第三处理模块，被配置为对

所述第一视频帧序列进行编码，以得到第一视频特征，还被配置为对所述第一分词序列进行编码，以得到第一分词特征；第四处理模块，被配置为对所述第二视频帧序列进行编码，以得到第二视频特征，还被配置为对所述第二分词序列进行编码，以得到第二分词特征；第五处理模块，被配置为利用所述第一视频特征、所述第一分词特征、
5 所述第二视频特征、所述第二分词特征确定预训练的目标函数；第六处理模块，被配置为利用所述预训练的目标函数进行多模态预训练。

根据本公开实施例的第三方面，提供一种多模态预训练装置，包括：存储器，被配置为存储指令；处理器，耦合到存储器，处理器被配置为基于存储器存储的指令执行实现如上述任一实施例所述的方法。

10 根据本公开实施例的第四方面，提供一种计算机可读存储介质，其中，计算机可读存储介质存储有计算机指令，指令被处理器执行时实现如上述任一实施例涉及的方法。

通过以下参照附图对本公开的示例性实施例的详细描述，本公开的其它特征及其优点将会变得清楚。

15

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性的前
20 提下，还可以根据这些附图获得其他的附图。

图 1 为本公开一个实施例的多模态预训练方法的流程示意图；

图 2 为本公开另一个实施例的多模态预训练方法的流程示意图；

图 3 为本公开一个实施例的多模态预训练装置的结构示意图；

图 4 为本公开另一个实施例的多模态预训练装置的结构示意图；

25 图 5 为本公开一个实施例的多模态预训练模型示意图。

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。

30 以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本公开及其

应用或使用的任何限制。基于本公开中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本公开保护的范围。

除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本公开的范围。

5 同时，应当明白，为了便于描述，附图中所示出的各个部分的尺寸并不是按照实际的比例关系绘制的。

对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为授权说明书的一部分。

在这里示出和讨论的所有示例中，任何具体值应被解释为仅仅是示例性的，而不是作为限制。因此，示例性实施例的其它示例可以具有不同的值。

10 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

发明人注意到，在相关技术中，视频文本多模态预训练技术为了挖掘两个模态之间的联系，仅利用带有掩码（Mask）的输入视频文本在预训练期间进行全局特征表示关联性的学习，这种学习方式使得输入视频帧和单词序列之间的整体视频-文本关系没有得到充分探索，从而导致多模态特征质量的下降。

据此，本公开提供一种多模态预训练方案，能够增强跨模态数据之间的关联性，有效提升多模态预训练模型对多模态数据内容的理解能力。

图 1 为本公开一个实施例的多模态预训练方法的流程示意图。在一些实施例中，
20 下列的多模态预训练方法由多模态预训练装置执行。

在步骤 101，对视频-文本对中的视频进行采样，以得到第一视频帧序列，对视频-文本对中的文本进行分词处理，以得到第一分词序列。

在一些实施例中，将视频按等间距采样的方式进行采样，以得到第一视频帧序列。

25 在一些实施例中，在第一分词序列的开头和结尾处分别设有标记[CLS]和[SEP]，以便后续处理方便。

在步骤 102，对第一视频帧序列进行掩码处理，以得到第二视频帧序列，对第一分词序列进行掩码处理，以得到第二分词序列。

在一些实施例中，用掩码以随机概率替换第一视频帧序列中的视频帧，以得到第二视频帧序列。

30 在一些实施例中，用掩码以随机概率替换第一分词序列中的分词，以得到第二分

词序列。

在步骤 103，对第一视频帧序列进行编码，以得到第一视频特征，对第一分词序列进行编码，以得到第一分词特征。

5 在一些实施例中，使用视频键值编码器（Video Key Encoder）对第一视频帧序列进行编码，以得到第一视频特征，使用文本键值编码器（Sentence Key Encoder）对第一分词序列进行编码，以得到第一分词特征。

视频键值编码器输出的第一视频特征反映了无掩码的视频帧的上下文特性。文本键值输出的第一分词特征反映了无掩码的分词序列的上下文特性。

由于视频键值和文本键值并不是本公开的发明点所在，因此这里不展开描述。

10 在步骤 104，对第二视频帧序列进行编码，以得到第二视频特征，对第二分词序列进行编码，以得到第二分词特征。

在一些实施例中，使用视频查询编码器（Video Query Encoder）对第二视频帧序列进行编码，以得到第二视频特征，使用文本查询编码器（Sentence Query Encoder）对第二分词序列进行编码，以得到第二分词特征。

15 视频查询编码器输出的第二视频特征反映了视频模态下帧与帧之间的关联性，文本查询编码器输出的第二分词特征反映了文本模态下词与词之间的关联性。

由于视频查询编码器和文本查询编码器并不是本公开的发明点所在，因此这里不展开描述。

20 在步骤 105，利用第一视频特征、第一分词特征、第二视频特征、第二分词特征确定预训练的目标函数。

在一些实施例中，确定预训练的目标函数如图 2 所示。

在步骤 201，利用第一分词特征、第二视频特征和预设的第一负样本特征确定第一对比损失值。

25 在一些实施例中，利用 MLP（Multi-layer Perceptron，多层感知机）模型将第一分词特征转换为全局的第一正样本特征 H_S^+ ，利用 MLP 模型将第二视频特征转换为全局的视频查询特征 H_V^m 。利用视频查询特征 H_V^m 、第一正样本特征 H_S^+ 和第一负样本特征 K_S^- 确定第一对比损失值。

需要说明的是，第一负样本特征 K_S^- 为：

$$K_S^- = \{H_{S,i}^-\}_{i=1}^K \quad (1)$$

30 其中 K 表示第一负样本特征所包括的负样本队列的大小， $H_{S,i}^-$ 表示负样本队列中

的第 i 个负样本。

在一些实施例中，利用公式（2）计算第一对比损失值 $L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-)$ 。

$$L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-) = -\log \frac{\exp(\langle H_V^m, H_S^+ \rangle / t)}{\exp(\langle H_V^m, H_S^+ \rangle / t) + \sum_{i=1}^K \exp(\langle H_V^m, H_{S,i}^- \rangle / t)} \quad (2)$$

其中， t 为用于控制缩放的超参数。运算符 $\langle \mathbf{A}, \mathbf{B} \rangle$ 表示向量 $\mathbf{A}$ 和 $\mathbf{B}$ 的余弦相似度。

- 5 在步骤 202，利用第一视频特征、第二分词特征和预设的第二负样本特征确定第二对比损失值。

在一些实施例中，利用 **MLP** 模型将第一视频特征转换为全局的第二正样本特征 H_V^+ ，利用 **MLP** 模型将第二分词特征转换为全局的文本查询特征 H_S^m 。利用文本查询特征 H_S^m 、第二正样本特征 H_V^+ 和第二负样本特征 K_V^- 确定第二对比损失值。

- 10 需要说明的是，第二负样本特征 K_V^- 为：

$$K_V^- = \{H_{V,i}^-\}_{i=1}^K \quad (3)$$

其中 K 表示第二负样本特征所包括的负样本队列的大小， $H_{V,i}^-$ 表示负样本队列中的第 i 个负样本。

在一些实施例中，利用公式（4）计算第二对比损失值 $L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-)$ 。

$$15 \quad L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-) = -\log \frac{\exp(\langle H_S^m, H_V^+ \rangle / t)}{\exp(\langle H_S^m, H_V^+ \rangle / t) + \sum_{i=1}^K \exp(\langle H_S^m, H_{V,i}^- \rangle / t)} \quad (4)$$

其中， t 为用于控制缩放的超参数。运算符 $\langle \mathbf{A}, \mathbf{B} \rangle$ 表示向量 $\mathbf{A}$ 和 $\mathbf{B}$ 的余弦相似度。

在步骤 203，根据第一对比损失值和第二对比损失值确定第一目标。

在一些实施例中，第一目标为第一对比损失值和第二对比损失值之和。例如，利用公式（5）计算第一目标。第一目标用于表示视频到文本和文本到视频的视频匹配损失

20 失的组合。

$$L_{Co-IM} = L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-) + L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-) \quad (5)$$

在步骤 204，利用第一视频特征、第二视频特征和第二负样本特征确定第三对比损失值。

- 25 在一些实施例中，利用视频查询特征 H_V^m 、第二正样本特征 H_V^+ 和第二负样本特征 K_V^- 确定第三对比损失值。

在一些实施例中，利用公式（6）计算第三对比损失值 $L_{NCE}^V(H_V^m, H_V^+, K_V^-)$ 。

$$L_{NCE}^V(H_V^m, H_V^+, K_V^-) = -\log \frac{\exp(\langle H_V^m, H_V^+ \rangle / t)}{\exp(\langle H_V^m, H_V^+ \rangle / t) + \sum_{i=1}^K \exp(\langle H_V^m, H_{V,i}^- \rangle / t)} \quad (6)$$

其中, t 为用于控制缩放的超参数。运算符 $\langle \mathbf{A}, \mathbf{B} \rangle$ 表示向量 $\mathbf{A}$ 和 $\mathbf{B}$ 的余弦相似度。

在步骤 205, 利用第一分词特征、第二分词特征和第一负样本特征确定第四对比损失值。

5 在一些实施例中, 利用文本查询特征 H_S^m 、第一正样本特征 H_S^+ 和第一负样本特征 K_S^- 确定第四对比损失值。

在一些实施例中, 利用公式 (7) 计算第四对比损失值 $L_{NCE}^S(H_S^m, H_S^+, K_S^-)$ 。

$$L_{NCE}^S(H_S^m, H_S^+, K_S^-) = -\log \frac{\exp(\langle H_S^m, H_S^+ \rangle / t)}{\exp\left(\frac{\langle H_S^m, H_S^+ \rangle}{t}\right) + \sum_{i=1}^K \exp(\langle H_S^m, H_{S,i}^- \rangle / t)} \quad (7)$$

其中, t 为用于控制缩放的超参数。运算符 $\langle \mathbf{A}, \mathbf{B} \rangle$ 表示向量 $\mathbf{A}$ 和 $\mathbf{B}$ 的余弦相似度。

在步骤 206, 根据第三对比损失值和第四对比损失值确定第二目标。

10 在一些实施例中, 第二目标为第三对比损失值和第四对比损失值之和。例如, 利用公式 (8) 计算第二目标。第二目标用于表示视频模态内和文本模态内的去噪损失。

$$L_{Co-ID} = L_{NCE}^V(H_V^m, H_V^+, K_V^-) + L_{NCE}^S(H_S^m, H_S^+, K_S^-) \quad (8)$$

在步骤 207, 根据第一目标和第二目标确定目标函数。

15 在一些实施例中, 目标函数为第一目标和第二目标之和。例如, 利用公式 (9) 计算目标函数 L 。

$$L = L_{Co-IM} + L_{Co-ID} \quad (9)$$

返回图 1。在步骤 106, 利用预训练的目标函数进行多模态预训练。

在本公开上述实施例提供的多模态预训练方法中, 基于跨模态匹配损失和模态内去噪损失来确定预训练的目标函数, 能够增强跨模态数据之间的关联性, 有效提升多模态预训练模型对多模态数据内容的理解能力。

20 在一些实施例中, 对第二视频特征和第二分词特征进行融合处理, 以得到融合特征。将融合特征输入 MLM (Masked Language Modelling, 带掩码的文本建模) 模型以得到第三目标 L_{MLM} , 将融合特征输入 MSG (Masked Language Generation, 带掩码的文本生成) 模型以得到第四目标 L_{MSG} 。

25 在一些实施例中, 使用跨模态解码器 (Cross-Modal Decoder) 对第二视频特征和第二分词特征进行融合处理, 以得到融合特征。跨模态解码器用于输出视频和文本多模态信息的融合特征, 为后续任务提供特征输入。

由于跨模态解码器并不是本公开的发明点所在, 因此这里不展开描述。

在一些实施例中, 根据第一目标 L_{Co-IM} 、第二目标 L_{Co-ID} 、第三目标 L_{MLM} 和第四

目标 L_{MSG} 确定目标函数 L 。

在一些实施例中，目标函数 L 为第一目标 L_{Co-IM} 、第二目标 L_{Co-ID} 、第三目标 L_{MLM} 和第四目标 L_{MSG} 之和。

例如，利用下列公式（10）计算目标函数 L 。

$$L = L_{Co-IM} + L_{Co-ID} + L_{MLM} + L_{MSG} \quad (10)$$

图3为本公开一个实施例的多模态预训练装置的结构示意图。如图3所示，多模态预训练装置包括第一处理模块31、第二处理模块32、第三处理模块33、第四处理模块34、第五处理模块35和第六处理模块36。

第一处理模块31被配置为对视频-文本对中的视频进行采样，以得到第一视频帧序列，还被配置为对视频-文本对中的文本进行分词处理，以得到第一分词序列。

在一些实施例中，将视频按等间距采样的方式进行采样，以得到第一视频帧序列。

在一些实施例中，在第一分词序列的开头和结尾处分别设有标记[CLS]和[SEP]，以便后续处理方便。

第二处理模块32被配置为对第一视频帧序列进行掩码处理，以得到第二视频帧序列，还被配置为对第一分词序列进行掩码处理，以得到第二分词序列。

在一些实施例中，用掩码以随机概率替换第一视频帧序列中的视频帧，以得到第二视频帧序列。

在一些实施例中，用掩码以随机概率替换第一分词序列中的分词，以得到第二分词序列。

第三处理模块33被配置为对第一视频帧序列进行编码，以得到第一视频特征，还被配置为对第一分词序列进行编码，以得到第一分词特征。

在一些实施例中，使用视频键值编码器对第一视频帧序列进行编码，以得到第一视频特征，使用文本键值编码器对第一分词序列进行编码，以得到第一分词特征。

视频键值编码器输出的第一视频特征反映了无掩码的视频帧的上下文特性。文本键值输出的第一分词特征反映了无掩码的分词序列的上下文特性。

第四处理模块34被配置为对第二视频帧序列进行编码，以得到第二视频特征，还被配置为对第二分词序列进行编码，以得到第二分词特征。

在一些实施例中，使用视频查询编码器对第二视频帧序列进行编码，以得到第二视频特征，使用文本查询编码器对第二分词序列进行编码，以得到第二分词特征。

视频查询编码器输出的第二视频特征反映了视频模态下帧与帧之间的关联性，文

本查询编码器输出的第二分词特征反映了文本模态下词与词之间的关联性。

第五处理模块 35 被配置为利用第一视频特征、第一分词特征、第二视频特征、第二分词特征确定预训练的目标函数。在一些实施例中，第五处理模块 35 利用第一分词特征、第二视频特征和预设的第一负样本特征确定第一对比损失值。

5 例如，利用 MLP 模型将第一分词特征转换为全局的第一正样本特征 H_S^+ ，利用 MLP 模型将第二视频特征转换为全局的视频查询特征 H_V^m 。利用视频查询特征 H_V^m 、第一正样本特征 H_S^+ 和第一负样本特征 K_S^- 确定第一对比损失值。

在一些实施例中，第一负样本特征 K_S^- 如上述公式 (1) 所示。

在一些实施例中，利用上述公式 (2) 计算第一对比损失值 $L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-)$ 。

10 第五处理模块 35 利用第一视频特征、第二分词特征和预设的第二负样本特征确定第二对比损失值。例如，利用 MLP 模型将第一视频特征转换为全局的第二正样本特征 H_V^+ ，利用 MLP 模型将第二分词特征转换为全局的文本查询特征 H_S^m 。利用文本查询特征 H_S^m 、第二正样本特征 H_V^+ 和第二负样本特征 K_V^- 确定第二对比损失值。

在一些实施例中，第二负样本特征 K_V^- 如上述公式 (3) 所示。

15 在一些实施例中，利用上述公式 (4) 计算第二对比损失值 $L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-)$ 。

第五处理模块 35 根据第一对比损失值和第二对比损失值确定第一目标。在一些实施例中，第一目标为第一对比损失值和第二对比损失值之和。例如，利用上述公式 (5) 计算第一目标。第一目标用于表示视频到文本和文本到视频的视频匹配损失的组合。

20 第五处理模块 35 利用第一视频特征、第二视频特征和第二负样本特征确定第三对比损失值。在一些实施例中，利用视频查询特征 H_V^m 、第二正样本特征 H_V^+ 和第二负样本特征 K_V^- 确定第三对比损失值。例如，利用上述公式 (6) 计算第三对比损失值 $L_{NCE}^V(H_V^m, H_V^+, K_V^-)$ 。

25 第五处理模块 35 利用第一分词特征、第二分词特征和第一负样本特征确定第四对比损失值。在一些实施例中，利用文本查询特征 H_S^m 、第一正样本特征 H_S^+ 和第一负样本特征 K_S^- 确定第四对比损失值。

在一些实施例中，利用上述公式 (7) 计算第四对比损失值 $L_{NCE}^S(H_S^m, H_S^+, K_S^-)$ 。

30 第五处理模块 35 根据第三对比损失值和第四对比损失值确定第二目标。在一些实施例中，第二目标为第三对比损失值和第四对比损失值之和。例如，利用上述公式 (8) 计算第二目标。第二目标用于表示视频模态内和文本模态内的去噪损失。

第五处理模块 35 根据第一目标和第二目标确定目标函数。在一些实施例中，目标函数为第一目标和第二目标之和。例如，利用上述公式（9）计算目标函数 L 。

在一些实施例中，第五处理模块 35 对第二视频特征和第二分词特征进行融合处理，以得到融合特征。将融合特征输入 MLM 模型以得到第三目标 L_{MLM} ，将融合特征
5 输入 MSG 模型以得到第四目标 L_{MSG} 。

在一些实施例中，使用跨模态解码器对第二视频特征和第二分词特征进行融合处理，以得到融合特征。跨模态解码器用于输出视频和文本多模态信息的融合特征，为后续任务提供特征输入。

在一些实施例中，根据第一目标 L_{Co-IM} 、第二目标 L_{Co-ID} 、第三目标 L_{MLM} 和第四
10 目标 L_{MSG} 确定目标函数 L 。在一些实施例中，目标函数 L 为第一目标 L_{Co-IM} 、第二目标 L_{Co-ID} 、第三目标 L_{MLM} 和第四目标 L_{MSG} 之和。例如，利用上述公式（10）计算目标函数 L 。

第六处理模块 36 被配置为利用预训练的目标函数进行多模态预训练。

图 4 为本公开另一个实施例的多模态预训练装置的结构示意图。如图 4 所示，多
15 模态预训练装置包括存储器 41 和处理器 42。

存储器 41 用于存储指令，处理器 42 耦合到存储器 41，处理器 42 被配置为基于存储器存储的指令执行实现如图 1 或图 2 中任一实施例涉及的方法。

如图 4 所示，该多模态预训练装置还包括通信接口 43，用于与其它设备进行信息交互。同时，该多模态预训练装置还包括总线 44，处理器 42、通信接口 43、以及存
20 储器 41 通过总线 44 完成相互间的通信。

存储器 41 可以包含高速 RAM 存储器，也可还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。存储器 41 也可以是存储器阵列。存储器 41 还可能被分块，并且块可按一定的规则组合成虚拟卷。

此外，处理器 42 可以是一个中央处理器 CPU，或者可以是专用集成电路 ASIC，
25 或是被配置成实施本公开实施例的一个或多个集成电路。

本公开同时还涉及一种计算机可读存储介质，其中计算机可读存储介质存储有计算机指令，指令被处理器执行时实现如图 1 或图 2 中任一实施例涉及的方法。

图 5 为本公开一个实施例的多模态预训练模型示意图。

如图 5 所示，通过对视频-文本对中的视频进行采样，以得到第一视频帧序列，对
30 视频-文本对中的文本进行分词处理，以得到第一分词序列。用掩码以随机概率替换第

一视频帧序列中的视频帧，以得到第二视频帧序列。用掩码以随机概率替换第一分词序列中的分词，以得到第二分词序列。

使用视频键值编码器对第一视频帧序列进行编码，以得到第一视频特征，使用文本键值编码器对第一分词序列进行编码，以得到第一分词特征。

- 5 使用视频查询编码器对第二视频帧序列进行编码，以得到第二视频特征，使用文本查询编码器对第二分词序列进行编码，以得到第二分词特征。

利用 **MLP** 模型将第一分词特征转换为全局的第一正样本特征 H_S^+ ，利用 **MLP** 模型将第一视频特征转换为全局的第二正样本特征 H_V^+ ，利用 **MLP** 模型将第二视频特征转换为全局的视频查询特征 H_V^m ，利用 **MLP** 模型将第二分词特征转换为全局的文本查询特征 H_S^m 。在 **Co-IM** (**Contrastive Inter-modal Matching**, 对比模态间匹配) 模块中，根据上述公式 (2)，利用视频查询特征 H_V^m 、第一正样本特征 H_S^+ 和第一负样本特征 K_S^- 确定第一对比损失值 $L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-)$ 。

在一些实施例中，第一负样本特征 K_S^- 如上述公式 (1) 所示。

- 15 根据上述公式 (4)，利用文本查询特征 H_S^m 、第二正样本特征 H_V^+ 和第二负样本特征 K_V^- 确定第二对比损失值 $L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-)$ 。

在一些实施例中，第二负样本特征 K_V^- 如上述公式 (3) 所示。

接下来，利用上述公式 (5) 计算第一目标 L_{Co-IM} 。

- 20 在 **Co-ID** (**Contrastive Intra-modal Denoising**, 对比模态内去噪) 模块中，根据上述公式 (6)，利用视频查询特征 H_V^m 、第二正样本特征 H_V^+ 和第二负样本特征 K_V^- 确定第三对比损失值 $L_{NCE}^V(H_V^m, H_V^+, K_V^-)$ 。

根据上述公式 (7)，利用文本查询特征 H_S^m 、第一正样本特征 H_S^+ 和第一负样本特征 K_S^- 确定第四对比损失值 $L_{NCE}^S(H_S^m, H_S^+, K_S^-)$ 。

接下来，根据上述公式 (8)，根据第三对比损失值和第四对比损失值确定第二目标 L_{Co-ID} 。

- 25 此外，使用跨模态解码器对第二视频特征和第二分词特征进行融合处理，以得到融合特征。将融合特征输入 **MLM** 模型以得到第三目标 L_{MLM} ，将融合特征输入 **MSG** 模型以得到第四目标 L_{MSG} 。

接下来，利用上述公式 (10)，通过将第一目标 L_{Co-IM} 、第二目标 L_{Co-ID} 、第三目标 L_{MLM} 和第四目标 L_{MSG} 之和作为目标函数 L 。

- 30 在一些实施例中，在上面所描述的功能单元模块可以实现为用于执行本公开所描

述功能的通用处理器、可编程逻辑控制器（**Programmable Logic Controller**，简称：**PLC**）、数字信号处理器（**Digital Signal Processor**，简称：**DSP**）、专用集成电路（**Application Specific Integrated Circuit**，简称：**ASIC**）、现场可编程门阵列（**Field-Programmable Gate Array**，简称：**FPGA**）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件或者其任意适当组合。

本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

本公开的描述是为了示例和描述起见而给出的，而并不是无遗漏的或者将本公开限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显然的。选择和描述实施例是为了更好说明本公开的原理和实际应用，并且使本领域的普通技术人员能够理解本公开从而设计适于特定用途的带有各种修改的各种实施例。

权 利 要 求

1. 一种多模态预训练方法，包括：

对视频-文本对中的视频进行采样，以得到第一视频帧序列；

对所述视频-文本对中的文本进行分词处理，以得到第一分词序列；

对所述第一视频帧序列进行掩码处理，以得到第二视频帧序列；

对所述第一分词序列进行掩码处理，以得到第二分词序列；

对所述第一视频帧序列进行编码，以得到第一视频特征，对所述第一分词序列进行编码，以得到第一分词特征；

对所述第二视频帧序列进行编码，以得到第二视频特征，对所述第二分词序列进行编码，以得到第二分词特征；

利用所述第一视频特征、所述第一分词特征、所述第二视频特征、所述第二分词特征确定预训练的目标函数；

利用所述预训练的目标函数进行多模态预训练。

2. 根据权利要求 1 所述的方法，其中，确定预训练的目标函数包括：

利用所述第一分词特征、所述第二视频特征和预设的第一负样本特征确定第一对比损失值；

利用所述第一视频特征、所述第二分词特征和预设的第二负样本特征确定第二对比损失值；

根据所述第一对比损失值和所述第二对比损失值确定第一目标；

利用所述第一视频特征、所述第二视频特征和所述第二负样本特征确定第三对比损失值；

利用所述第一分词特征、所述第二分词特征和所述第一负样本特征确定第四对比损失值；

根据所述第三对比损失值和所述第四对比损失值确定第二目标；

根据所述第一目标和所述第二目标确定所述目标函数。

3. 根据权利要求 2 所述的方法，其中，确定第一对比损失值包括：

将所述第一分词特征转换为全局的第一正样本特征；

将所述第二视频特征转换为全局的视频查询特征；

利用所述视频查询特征、所述第一正样本特征和所述第一负样本特征确定第一对比损失值。

4. 根据权利要求 3 所述的方法，其中，确定第二对比损失值包括：

将所述第一视频特征转换为全局的第二正样本特征；

将所述第二分词特征转换为全局的文本查询特征；

利用所述文本查询特征、所述第二正样本特征和所述第二负样本特征确定第二对比损失值。

5. 根据权利要求 4 所述的方法，其中，确定第三对比损失值包括：

利用所述视频查询特征、所述第二正样本特征和所述第二负样本特征确定第三对比损失值。

6. 根据权利要求 5 所述的方法，其中，确定第四对比损失值包括：

利用所述文本查询特征、所述第一正样本特征和所述第一负样本特征确定第四对比损失值。

7. 根据权利要求 2 所述的方法，其中，

所述第一目标为所述第一对比损失值和所述第二对比损失值之和；

所述第二目标为所述第三对比损失值和所述第四对比损失值之和。

8. 根据权利要求 2-7 中任一项所述的方法，其中，

所述目标函数为所述第一目标和所述第二目标之和。

9. 根据权利要求 2-7 中任一项所述的方法，还包括：

对所述第二视频特征和所述第二分词特征进行融合处理，以得到融合特征；

将所述融合特征输入带掩码的文本建模 MLM 模型以得到第三目标，将所述融合特征输入带掩码的文本生成 MSG 模型以得到第四目标；

所述根据所述第一目标和所述第二目标确定所述目标函数包括：

根据所述第一目标、所述第二目标、所述第三目标和所述第四目标确定所述目标函数。

10. 根据权利要求 9 所述的方法，其中，
所述目标函数为所述第一目标、所述第二目标、所述第三目标和所述第四目标之和。

11. 一种多模态预训练装置，包括：

第一处理模块，被配置为对视频-文本对中的视频进行采样，以得到第一视频帧序列，还被配置为对所述视频-文本对中的文本进行分词处理，以得到第一分词序列；

第二处理模块，被配置为对所述第一视频帧序列进行掩码处理，以得到第二视频帧序列，还被配置为对所述第一分词序列进行掩码处理，以得到第二分词序列；

第三处理模块，被配置为对所述第一视频帧序列进行编码，以得到第一视频特征，还被配置为对所述第一分词序列进行编码，以得到第一分词特征；

第四处理模块，被配置为对所述第二视频帧序列进行编码，以得到第二视频特征，还被配置为对所述第二分词序列进行编码，以得到第二分词特征；

第五处理模块，被配置为利用所述第一视频特征、所述第一分词特征、所述第二视频特征、所述第二分词特征确定预训练的目标函数；

第六处理模块，被配置为利用所述预训练的目标函数进行多模态预训练。

12. 一种多模态预训练装置，包括：

存储器，被配置为存储指令；

处理器，耦合到存储器，处理器被配置为基于存储器存储的指令执行实现如权利要求 1-10 中任一项所述的方法。

13. 一种非瞬态计算机可读存储介质，其中，计算机可读存储介质存储有计算机指令，指令被处理器执行时实现如权利要求 1-10 中任一项所述的方法。