

## 用于生成字幕器以及输出字幕的方法和装置

### 相关申请的交叉引用

本公开要求于 2021 年 3 月 30 日提交的申请号为 202110338045.X、发明名称为“用于生成字幕器以及输出字幕的方法和装置”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

本公开的实施例涉及计算机技术领域，具体涉及用于生成字幕器以及输出字幕的方法和装置。

### 背景技术

图像字幕是一个新兴且发展迅速的研究主题，它是一种用自然语言句子自动描述图像的技术。

相关技术中，大部分在带注释的图像-句子对上训练字幕器，他们都遵循先利用卷积神经网络对输入图像进行编码，然后利用循环神经网络对句子进行解码的编解码器范式。一系列的工作都在升级图像字幕的注意机制，以增强视觉内容和自然句子之间的跨域基础。

相关技术严重依赖于大量的训练图像句子对，一方面，这些训练图像句子对的获取是极为昂贵且耗时的。另一方面，过分依赖训练图像句子对，阻碍了字幕器的广泛应用。

### 发明内容

本公开的实施例提出了用于生成字幕器的方法和装置以及用于输出字幕的方法和装置。

本公开的实施例提供了一种用于生成字幕器的方法，包括：获取样本图像集；将样本图像集输入句子生成器的图像编码器，输出对象集；将对象集分组成第一对象集和第二对象集，其中，第一对象集为被包含在预定对象集内的对象集，第二对象集为被排除在预定对象集外的对象集；将图像编码器输出的对象集输入句子生成器的句子解码器，在解码步骤中以第一对象集、第二对象集为约束条件进行波束搜索，生成伪图像句子对集；将伪图像句子

对集作为样本集训练句子生成器，得到字幕器。

在一些实施例中，该方法还包括：通过以下至少一种方式优化字幕器：通过句子鉴别器对字幕器进行对抗式训练来优化字幕器；通过字幕器识别出的对象在字幕器输出的句子中的包含程度优化字幕器；通过图像三元组与相应生成的句子之间的语义相关性优化字幕器，其中，图像三元组包括查询图像，正图像和负图像。

在一些实施例中，通过句子鉴别器对字幕器进行对抗式训练来优化字幕器，包括：提取预置的第一样本集，其中，每个第一样本包括图像和对应的真句子；提取预先建立的生成式对抗网络，其中，生成式对抗网络包括字幕器和句子鉴别器，字幕器用于对所输入的图像进行图像编码后再进行句子解码，得到伪句子，句子鉴别器用于确定所输入的句子是否为字幕器所输出的伪句子；基于机器学习方法，从第一样本集中选取第一样本，以及执行以下第一训练步骤：将选取的第一样本中的图像输入字幕器，输出伪句子；将伪句子和选取的第一样本中的真句子输入句子鉴别器，输入鉴别结果；根据输出的鉴别结果统计句子鉴别器的准确率；若准确率达到预设数值，则确定出字幕器训练完成。

在一些实施例中，该方法还包括：若准确率未达到预设数值，则计算句子鉴别器的对抗性损失，调整句子鉴别器的相关参数使得对抗性损失减小，以及从第一样本集中重新选取第一样本，继续执行第一训练步骤。

在一些实施例中，该方法还包括：若准确率未达到预设数值，则计算字幕器的对抗性奖励，调整字幕器的相关参数使得对抗性奖励增大，以及从第一样本集中重新选取第一样本，继续执行第一训练步骤。

在一些实施例中，通过字幕器识别出的对象在字幕器输出的句子中的包含程度优化字幕器，包括：提取预置的第二样本集，其中，每个第二样本包括图像；基于机器学习方法，从第二样本集中选取样本，以及执行以下第二训练步骤：将选取的第二样本中的图像输入字幕器的图像编码器，输出样本对象集；将样本对象集输入字幕器的句子解码器，输出伪句子；计算伪句子中包含样本对象集中的样本对象的置信度均值分数，作为伪句子的对象包含奖励；若对象包含奖励达到预设包含奖励阈值，则确定出字幕器训练完成。

在一些实施例中，该方法还包括：若对象包含奖励未达到预设包含奖励阈值，则调整字幕器的相关参数使得对象包含奖励增大，以及从第二样本集

中重新选取第二样本，继续执行第二训练步骤。

在一些实施例中，通过图像三元组与相应生成的句子之间的语义相关性优化字幕器，包括：提取预置的第三样本集，其中，每个第三样本包括查询图像、正图像和负图像，正图像与查询图像共享至少两个对象，而负图像和查询图像没有任何共同的对象；基于机器学习方法，从第三样本集中选取第三样本，以及执行以下第三训练步骤：将选取的第三样本中的查询图像、正图像和负图像分别输入字幕器，输出查询句子、正句子和负句子；计算查询句子和正句子的第一语义相似度以及计算查询句子和负句子的第二语义相似度；根据第一语义相似度和第二语义相似度计算自监督三重态损失；若自监督三重态损失小于预定损失阈值，则确定出字幕器训练完成。

在一些实施例中，该方法还包括：若自监督三重态损失不小于预定损失阈值，则调整字幕器的相关参数使得自监督三重态损失减小，以及从第三样本集中重新选取第三样本，继续执行第三训练步骤。

在一些实施例中，计算查询句子和正句子的第一语义相似度以及计算查询句子和负句子的第二语义相似度，包括：对于查询句子、正句子和负句子，分别计算句子中每一个词的基于对象的概率分布，进行最大池化操作，分别得到查询句子特征、正句子特征和负句子特征；计算查询句子特征和正句子特征的第一语义相似度以及计算查询句子特征和负句子特征的第二语义相似度。

在一些实施例中，该方法还包括：若对抗性奖励、对象包含奖励和自监督三重态损失的加权和大于预定目标值，则调整字幕器的相关参数使得加权和减小。

在一些实施例中，图像编码器包括具有区域级别注意机制的两层 LSTM，其中，第一层 LSTM 充当自上而下的注意模块，根据上下文信息计算对象级别的注意，而第二层 LSTM 是用于生成句子的语言模型。

本公开的实施例还提供了一种用于输出字幕的方法，包括：获取待处理的图像；将图像输入采用如上所述的用于生成字幕器的方法生成的字幕器中，输出图像对应的字幕。

本公开的实施例还提供了一种用于生成字幕器的装置，包括：获取单元，被配置成获取样本图像集；编码单元，被配置成将样本图像集输入句子生成器的图像编码器，输出对象集；分组单元，被配置成将对象集分组成第一对

象集和第二对象集，其中，第一对象集为被包含在预定对象集内的对象集，第二对象集为被排除在预定对象集外的对象集；解码单元，被配置成将图像编码器输出的对象集输入句子生成器的句子解码器，在解码步骤中以第一对象集、第二对象集为约束条件进行波束搜索，生成伪图像句子对集；训练单元，被配置成将伪图像句子对集作为样本集训练句子生成器，得到字幕器。

在一些实施例中，该装置还包括优化单元，被配置成：通过以下至少一种方式优化字幕器：通过句子鉴别器对字幕器进行对抗式训练来优化字幕器；通过字幕器识别出的对象在字幕器输出的句子中的包含程度优化字幕器；通过图像三元组与相应生成的句子之间的语义相关性优化字幕器，其中，图像三元组包括查询图像，正图像和负图像。

在一些实施例中，优化单元进一步被配置成：提取预置的第一样本集，其中，每个第一样本包括图像和对应的真句子；提取预先建立的生成式对抗网络，其中，生成式对抗网络包括字幕器和句子鉴别器，字幕器用于对所输入的图像进行图像编码后再进行句子解码，得到伪句子，句子鉴别器用于确定所输入的句子是否为字幕器所输出的伪句子；基于机器学习装置，从第一样本集中选取第一样本，以及执行以下第一训练步骤：将选取的第一样本中的图像输入字幕器，输出伪句子；将伪句子和选取的第一样本中的真句子输入句子鉴别器，输入鉴别结果；根据输出的鉴别结果统计句子鉴别器的准确率；若准确率达到预设数值，则确定出字幕器训练完成。

在一些实施例中，优化单元进一步被配置成：若准确率未达到预设数值，则计算句子鉴别器的对抗性损失，调整句子鉴别器的相关参数使得对抗性损失减小，以及从第一样本集中重新选取第一样本，继续执行第一训练步骤。

在一些实施例中，优化单元进一步被配置成：若准确率未达到预设数值，则计算字幕器的对抗性奖励，调整字幕器的相关参数使得对抗性奖励增大，以及从第一样本集中重新选取第一样本，继续执行第一训练步骤。

在一些实施例中，优化单元进一步被配置成：提取预置的第二样本集，其中，每个第二样本包括图像；基于机器学习装置，从第二样本集中选取样本，以及执行以下第二训练步骤：将选取的第二样本中的图像输入字幕器的图像编码器，输出样本对象集；将样本对象集输入字幕器的句子解码器，输出伪句子；计算伪句子中包含样本对象集中的样本对象的置信度均值分数，作为伪句子的对象包含奖励；若对象包含奖励达到预设包含奖励阈值，则确

定出字幕器训练完成。

在一些实施例中，优化单元进一步被配置成：若对象包含奖励未达到预设包含奖励阈值，则调整字幕器的相关参数使得对象包含奖励增大，以及从第二样本集中重新选取第二样本，继续执行第二训练步骤。

在一些实施例中，优化单元进一步被配置成：提取预置的第三样本集，其中，每个第三样本包括查询图像、正图像和负图像，正图像与查询图像共享至少两个对象，而负图像和查询图像没有任何共同的对象；基于机器学习装置，从第三样本集中选取第三样本，以及执行以下第三训练步骤：将选取的第三样本中的查询图像、正图像和负图像分别输入字幕器，输出查询句子、正句子和负句子；计算查询句子和正句子的第一语义相似度以及计算查询句子和负句子的第二语义相似度；根据第一语义相似度和第二语义相似度计算自监督三重态损失；若自监督三重态损失小于预定损失阈值，则确定出字幕器训练完成。

在一些实施例中，优化单元进一步被配置成：若自监督三重态损失不小于预定损失阈值，则调整字幕器的相关参数使得自监督三重态损失减小，以及从第三样本集中重新选取第三样本，继续执行第三训练步骤。

在一些实施例中，优化单元进一步被配置成：对于查询句子、正句子和负句子，分别计算句子中每一个词的基于对象的概率分布，进行最大池化操作，分别得到查询句子特征、正句子特征和负句子特征；计算查询句子特征和正句子特征的第一语义相似度以及计算查询句子特征和负句子特征的第二语义相似度。

在一些实施例中，优化单元进一步被配置成：若对抗性奖励、对象包含奖励和自监督三重态损失的加权和大于预定目标值，则调整字幕器的相关参数使得加权和减小。

在一些实施例中，图像编码器包括具有区域级别注意机制的两层 LSTM，其中，第一层 LSTM 充当自上而下的注意模块，根据上下文信息计算对象级别的注意，而第二层 LSTM 是用于生成句子的语言模型。

本公开的实施例提供了一种用于输出字幕的装置，包括：获取单元，被配置成获取待处理的图像；输出单元，被配置成将图像输入采用如上所述的用于生成字幕器的方法生成的字幕器中，输出图像对应的字幕。

本公开的实施例提供了一种电子设备，包括：一个或多个处理器；存储

装置，其上存储有一个或多个计算机程序，当一个或多个计算机程序被一个或多个处理器执行，使得一个或多个处理器实现如上所述的用于生成字幕器的方法。

本公开的实施例提供了一种计算机可读介质，其上存储有计算机程序，其中，计算机程序被处理器执行时实现如上所述的用于生成字幕器的方法。

#### 附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本公开的其它特征、目的和优点将会变得更明显：

图 1 是本公开的一个实施例可以应用于其中的示例性系统架构图；

图 2 是根据本公开的用于生成字幕器的方法的一个实施例的流程图；

图 3 是根据本公开的用于生成字幕器的方法的一个应用场景的示意图；

图 4 是根据本公开的用于输出字幕的方法的一个实施例的流程图；

图 5 是根据本公开的用于生成字幕器的装置的一个实施例的结构示意图；

图 6 是根据本公开的用于输出字幕的装置的一个实施例的结构示意图；

图 7 是适于用来实现本公开的实施例的电子设备的计算机系统的结构示意图。

#### 具体实施方式

下面结合附图和实施例对本公开作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关方案，而非对该方案的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关方案相关的部分。

需要说明的是，在不冲突的情况下，本公开中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本公开。

图 1 示出了可以应用本公开实施例的用于生成字幕器的方法、用于生成字幕器的装置、用于输出字幕的方法或用于输出字幕的装置的示例性系统架构 100。

如图 1 所示，系统架构 100 可以包括终端 101、102，网络 103、数据库服务器 104 和服务器 105。网络 103 用以在终端 101、102，数据库服务器 104 与服务器 105 之间提供通信链路的介质。网络 103 可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

用户 110 可以使用终端 101、102 通过网络 103 与服务器 105 进行交互，以接收或发送消息等。终端 101、102 上可以安装有各种客户端应用，例如模型训练类应用、字幕生成类应用、图像处理类应用、购物类应用、支付类应用、网页浏览器和即时通讯工具等。

这里的终端 101、102 可以是硬件，也可以是软件。当终端 101、102 为硬件时，可以是具有显示屏的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、MP3 播放器 (Moving Picture Experts Group Audio Layer III, 动态影像专家压缩标准音频层面 3)、膝上型便携计算机和台式计算机等等。当终端 101、102 为软件时，可以安装在上述所列举的电子设备中。其可以实现成多个软件或软件模块 (例如用来提供分布式服务)，也可以实现成单个软件或软件模块。在此不做具体限定。

当终端 101、102 为硬件时，其上还可以安装有图像采集设备。图像采集设备可以是各种能实现采集图像功能的设备，如摄像头、传感器等等。用户 110 可以利用终端 101、102 上的图像采集设备，来采集各种场景的图像。

数据库服务器 104 可以是提供各种服务的数据库服务器。例如数据库服务器中可以存储有样本集。样本集中包含有大量的样本。其中，样本可以包括样本图像以及与样本图像对应的句子。这样，用户 110 也可以通过终端 101、102，从数据库服务器 104 所存储的样本集中选取样本。

服务器 105 也可以是提供各种服务的服务器，例如对终端 101、102 上显示的各种应用提供支持的后台服务器。后台服务器可以利用终端 101、102 发送的样本集中的样本，对初始模型进行训练，并可以将训练结果 (如生成的字幕器) 发送给终端 101、102。这样，用户可以应用生成的字幕器为图像生成字幕。

这里的数据库服务器 104 和服务器 105 同样可以是硬件，也可以是软件。当它们为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当它们为软件时，可以实现成多个软件或软件模块 (例如用来提供分布式服务)，也可以实现成单个软件或软件模块。在此不做具体限定。

需要说明的是，本公开实施例所提供的用于生成字幕器的方法或用于输出字幕的方法一般由服务器 105 执行。相应地，用于生成字幕器的装置或用于输出字幕的装置一般也设置于服务器 105 中。

需要指出的是，在服务器 105 可以实现数据库服务器 104 的相关功能的情况下，系统架构 100 中可以不设置数据库服务器 104。

应该理解，图 1 中的终端、网络、数据库服务器和服务器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终端、网络、数据库服务器和服务器。

继续参见图 2，其示出了根据本公开的用于生成字幕器的方法的一个实施例的流程 200。该用于生成字幕器的方法可以包括以下步骤：

步骤 201，获取样本图像集。

在本实施例中，用于生成字幕器的方法的执行主体（例如图 1 所示的服务器）可以从数据库服务器获取预先存储的样本图像集。也可从终端获取终端拍摄的图像作为样本图像。

步骤 202，将样本图像集输入句子生成器的图像编码器，输出对象集。

在本实施例中，句子生成器是初始的字幕器，是一种将输入的图像转换成句子的神经网络。句子生成器可包括图像编码器和句子解码器。

图像编码器为每个输入图像生成中间表示，本公开以最常见的对象检测模型（Faster R-CNN）作为图像编码器来检测图像中的对象，实际应用中还可使用其它的图像编码器。将每个图像  $I_i$  编码成一组显著图像区域  $V_i = \{v_k\}_{k=1}^K$ ，其中包含 K 个检测到的对象，例如，人、花、草、树、椅子、狗等。

步骤 203，将对象集分组成第一对象集和第二对象集。

在本实施例中，第一对象集为被包含在预定对象集内的对象集，第二对象集为被排除在预定对象集外的对象集。从技术上讲，给定输入图像  $I_i$ ，通过对象检测模型（例如 Faster R-CNN）输出识别对象集  $\{O_k, c_{(O_k)}\}_{k=1}^K$ ，其中  $O_k$  是第 K 个图像区域中具有最高置信度得分的被识别对象， $c_{(O_k)}$  是相应的置信度得分。以 COCO 数据集中最常见的 80 个检测对象为基准，将识别到的对象重新分组成需要被包含在内的对象集  $O_R = \{O_r^R\}_{r=1}^{K_R}$  和被排除在外的对象集  $O_E = \{O_e^E\}_{e=1}^{K_E}$ 。例如，预定对象集内包括房屋、汽车、人、花、草、树，而对象集包括人、花、草、树、椅子、狗，则第一对象集（被包含在预定对象集内的对象集）包括人、花、草、树，而第二对象集（被排除在外的对象集）包括椅子、狗。

步骤 204，将图像编码器输出的对象集输入句子生成器的句子解码器，在解码步骤中以第一对象集、第二对象集为约束条件进行波束搜索，生成伪图像句子对集。

在本实施例中，给定图像编码器产生的中间表示，利用句子解码器对输

出的句子逐字进行解码。参考自下而上的注意力模型 (Bottom-up and Top-Down), 可将句子解码器实现为具有区域级别注意机制的两层 LSTM (Long Short-Term Memory, 长短期记忆网络)。第一层 LSTM ( $LSTM^1$ ) 充当自上而下的注意模块, 根据上下文信息计算对象级别的注意, 而第二层 LSTM ( $LSTM^2$ ) 是用于生成句子的语言模型。在每个解码步骤  $t$  处, 第二层 LSTM 的隐藏状态  $h_{t-1}^2$ 、编码图像特征的均值  $\bar{v} = \frac{1}{K} \sum_{k=1}^K v_k$  以及输入字  $w_{t-1}$  被视为上下文信息, 并馈入第一层 LSTM, 由此得到第一层 LSTM 的隐藏状态:

$$h_t^1 = LSTM^1([h_{t-1}^2, \bar{v}, W_\mu w_{t-1}], h_{t-1}^1)$$

其中  $W_\mu$  是单词嵌入矩阵,  $w_{t-1}$  是词编码, 基于隐藏状态  $h_t^1$  测量所有  $K$  个图像区域的注意力分布如下:

$$\begin{aligned} a_{t,k} &= w_a^T \tanh(W_{va} v_k + W_{ha} h_t^1) \\ \lambda_t &= \text{softmax}(a_t) \end{aligned}$$

其中  $a_{t,k}$  表示  $a_t$  的第  $K$  个元素,  $W_{va}$ ,  $W_{ha}$ ,  $w_a^T$  是变换矩阵。利用注意力分布加权得到图像特征:

$$\hat{v}_t = \sum_{k=1}^K \lambda_{t,k} v_k$$

其中  $\lambda_{t,k}$  是  $\lambda_t$  的第  $k$  个元素, 代表了图像区域  $v_k$  的注意力概率。接着串联图像特征  $\hat{v}_t$  和隐藏状态  $h_t^1$  做为第二层 LSTM 的输入, 得到第二层 LSTM 的隐藏状态:

$$\begin{aligned} h_t^2 &= LSTM^2([\hat{v}_t, h_t^1], h_{t-1}^2) \\ p(w_t) &= \text{softmax}(W_E h_t^2) \end{aligned}$$

其中  $W_E$  是一个线性嵌入矩阵, 它将  $h_t^2$  投影到词汇空间以进行单词预测。

通过预训练的字幕器生成伪图像句子对, 一种自然的方法是采用波束搜索 (beam search), 它是一种启发式搜索算法, 它在每个解码步骤保持波束  $B_t$ , 包含  $b$  个最可能的部分句子。但是, 输入图像和输出句子之间的语义相关性并未完全用于推理时的句子生成。为了缓解这个问题, 本公开设计了语义约束波束搜索, 重新构造了波束搜索以确保包含识别的对象并排除无关的对象。

从技术上讲, 给定输入图像  $I_i$ , 通过对象检测模型 (例如 FasterR-CNN) 输出识别对象集  $\{O_k, c_{(o_k)}\}_{k=1}^K$ , 其中  $O_k$  是第  $K$  个图像区域中具有最高置信度得分的被识别对象,  $c_{(o_k)}$  是相应的置信度得分。以 COCO 数据集中最常见的 80 个检测对象为基准, 将识别到的对象重新分组成为需要被包含在内的对象集

$\mathcal{O}_R = \{\mathcal{O}_r^R\}_{r=1}^{K_R}$  和被排除在外的对象集  $\mathcal{O}_E = \{\mathcal{O}_e^E\}_{e=1}^{K_E}$ 。将识别到  $\mathcal{O}_R$  中的对象和排除掉  $\mathcal{O}_E$  中的对象作为一种约束条件，并且用有限状态机 (Finite-state machine) 来执行这种约束，这样有限状态机就可以识别出满足所有对象包含约束的词序列，然后将波束搜索算法与有限状态机相结合。具体地说，在有限状态机中为每个状态  $a \in A$  保持搜索波束  $B_t^a$ ，在每个解码步骤  $t$ ，通过在候选集  $E_t^a$  中保留  $b$  个最可能的部分字序列来更新每个波束  $B_t^a$ ：

$$E_t^a = \bigcup_{a' \in A} \{(w_{1:t-1}, w_t) | w_{1:t-1} \in B_{t-1}^{a'}, w_t \in V - \mathcal{O}_E, \delta(a', w_t) = a\},$$

其中  $w_{1:t-1}$  表示输出句子长度为  $t-1$ ， $V$  是词典， $\delta: A \times \{V - \mathcal{O}_E\} \rightarrow A$  是有限状态机中的状态转移函数。这里只从词汇表中查找单词（没有  $\mathcal{O}_E$  中的无关对象），以扩展当前的部分单词序列。因此，有限状态机的设计要求接受状态的词序列必须满足包含条件，同时排除所有无关对象。

伪图像句子对集中每个伪图像句子对包括图像和句子，图像和句子可以是非配对的。

步骤 205，将伪图像句子对集作为样本集训练句子生成器，得到字幕器。

在本实施例中，可用  $P = \{I_i, S_i^*\}_{i=1}^{N_I}$  表示伪图像句子对的集合，其中  $S_i^* = \{w_t^*\}_{t=1}^T$  代表产生的伪句子，利用这些伪图像句子对，可以直接训练一个具有以下交叉熵损失的字幕器：

$$\mathcal{L}_{XE} = - \sum_{t=1}^T \log(p_{\theta}(w_t^* | w_0^*, \dots, w_{t-1}^*))$$

这里  $\theta$  表示句子解码器中的参数。

在本实施例的一些可选的实现方式中，该方法还包括：通过以下至少一种方式优化字幕器：通过句子鉴别器对字幕器进行对抗式训练来优化字幕器；通过字幕器识别出的对象在字幕器输出的句子中的包含程度优化字幕器；通过图像三元组与相应生成的句子之间的语义相关性优化字幕器，其中，图像三元组包括查询图像，正图像和负图像。

可通过上述任一方式优化字幕器，也可任意两种方式的组合进行优化。还可将三种方式结合在一起优化字幕器。

在本实施例的一些可选的实现方式中，通过句子鉴别器对所述字幕器进行对抗式训练来优化所述字幕器，包括：提取预置的第一样本集，其中，每个第一样本包括图像和对应的真句子；提取预先建立的生成式对抗网络，其

中, 所述生成式对抗网络包括字幕器和句子鉴别器, 所述字幕器用于对所输入的图像进行图像编码后再进行句子解码, 得到伪句子, 所述句子鉴别器用于确定所输入的句子是否为所述字幕器所输出的伪句子; 基于机器学习方法, 从所述第一样本集中选取第一样本, 以及执行以下第一训练步骤: 将选取的第一样本中的图像输入所述字幕器, 输出伪句子; 将所述伪句子和选取的第一样本中的真句子输入所述句子鉴别器, 输入鉴别结果; 根据输出的鉴别结果统计所述句子鉴别器的准确率; 若所述准确率达到预设数值, 则确定出所述字幕器训练完成。

若所述准确率未达到预设数值, 则计算所述句子鉴别器的对抗性损失, 调整所述句子鉴别器的相关参数使得所述对抗性损失减小, 以及从所述第一样本集中重新选取第一样本, 继续执行所述第一训练步骤。

若所述准确率未达到预设数值, 则计算所述字幕器的对抗性奖励, 调整所述字幕器的相关参数使得所述对抗性奖励增大, 以及从所述第一样本集中重新选取第一样本, 继续执行所述第一训练步骤。

句子鉴别器的结构如图3所示。句子鉴别器和字幕器(包括图像编码器和句子解码器)组成了生成式对抗网络。句子鉴别器用于区分输入句子是未配对句子数据集中的真实句子还是由字幕器生成的伪句子。基于递归神经网络(RNN)的句子建模, 可利用LSTM在上下文上将单词序列编码为句子级别的表示形式, 以识别真实/生成的句子。从技术上讲, 给定包含T个单词的句子 $S = \{w_t\}_{t=1}^T$ , 句子鉴别器中的LSTM以自然顺序读取输入的单词序列:

$$h_t^d = \text{LSTM}^d(W_\gamma w_{t-1}, h_{t-1}^d), t \in \{1 \dots T\}$$

其中 $h_t^d$ 是t时刻的输出隐藏状态,  $W_\gamma \in \mathbb{R}^{D_E \times D_d}$ 是词嵌入矩阵。之后, 将最终的输出隐藏状态 $h_T^d$ 作为句子级别的表示。利用 $h_T^d$ 通过以下sigmoid函数 $D(S)$ 来产生识别真实句子性的概率:

$$D(S) = \text{sigmoid}(W_{FC}(h_T^d))$$

其中 $W_{FC}$ 是全连层的嵌入矩阵。

每次训练过程中, 句子鉴别器判断输入的句子是真句子还是字幕器生成的伪句子。统计鉴别结果是否正确, 如果准确率达到预设数值(例如0.5), 则说明字幕器的伪造句子的效果很好, 能够骗得过句子鉴别器, 则结束训练。否则需要调整字幕器和句子鉴别器的网络参数, 再进行训练。可先固定字幕器的参数, 调整句子鉴别器的参数进行训练, 然后固定句子鉴别器的参数,

调整字幕器的参数进行训练。交替调整句子鉴别器和字幕器的参数，最后训练完成了句子鉴别器和字幕器。而实际应用的是字幕器。

对抗性奖励 (Adversarial Reward)，为了生成与人类书写的字幕不可区分的句子，本公开采用对抗式训练和句子级别的对抗式奖励，使生成的句子分布与手动描述的句子分布相匹配。从技术上讲，图像字幕器 $G$ 被视为句子生成器，捕获数据分布以生成句子。句子鉴别器 $D$ 以从实际句子或 $G$ 生成的句子中随机选择的句子作为输入，并且在两个句子源（即生成的句子或真实的句子）上产生概率分布 $D(S)$ 。通过对抗性训练之后，在双人博弈中训练图像字幕器和句子鉴别器。具体来说，句子鉴别器 $D$ 通过正确区分真实句子 $\{S_i\}$ 和生成句子 $\{\hat{S}_i\}$ 来优化，即最小化对抗性损失：

$$\mathcal{L}_{adv} = -\log(D(S_i)) - \log(1 - D(\hat{S}_i))$$

同时，图像字幕器 $G$ 通过最大化对抗性奖励 $r_{adv}$ 来学习，旨在用生成的句子愚弄句子鉴别器：

$$r_{adv}(\hat{S}_i) = \log(D(\hat{S}_i))$$

通过生成式对抗网络可以提高字幕器的准确率。

在本实施例的一些可选的实现方式中，通过所述字幕器识别出的对象在所述字幕器输出的句子中的包含程度优化所述字幕器，包括：

提取预置的第二样本集，其中，每个第二样本包括图像；

基于机器学习方法，从所述第二样本集中选取样本，以及执行以下第二训练步骤：将选取的第二样本中的图像输入所述字幕器的图像编码器，输出样本对象集；将所述样本对象集输入字幕器的句子解码器，输出伪句子；计算所述伪句子中包含所述样本对象集中的样本对象的置信度均值分数，作为所述伪句子的对象包含奖励；若所述对象包含奖励达到预设包含奖励阈值，则确定出所述字幕器训练完成。

若所述对象包含奖励未达到预设包含奖励阈值，则调整所述字幕器的相关参数使得所述对象包含奖励增大，以及从所述第二样本集中重新选取第二样本，继续执行所述第二训练步骤。

对象包含奖励 (Object Inclusion Reward)，由于对抗性奖励只强化字幕器以产生更逼真的句子，而没有明确的概念来描述图像内容和生成句子之间的语义关联。因此，本公开进一步将识别出的目标在输出句子中的包含程度作为一个额外的自监督目标，即对象包含奖励，以鼓励字幕器描述生成句子包

含识别出的目标。通过这种方式，强调了两者之间的语义相关性，加强了生成的字幕的质量。具体地，本公开用所有识别出的对象构建一个包含对象集合  $\mathcal{O}_i = \{\mathbf{o}_k\}_{k=1}^K$ ，给定生成地句子  $\hat{\mathbf{S}}_i = \{\hat{\mathbf{w}}_t\}_{t=1}^T$ ，通过计算句子中包含的在上述集合  $\mathcal{O}_i$  中的对象的置信度均值分数来构建对象包含奖励：

$$I(\hat{\mathbf{w}}_t) = \begin{cases} \eta, & \text{if } \hat{\mathbf{w}}_t \in \mathcal{O}_i \\ 0, & \text{if } \hat{\mathbf{w}}_t \notin \mathcal{O}_i \end{cases}$$

$$r_{obj}(\hat{\mathbf{S}}_i) = \sum_{t=1}^T I(\hat{\mathbf{w}}_t) c(\hat{\mathbf{w}}_t)$$

其中  $I(\hat{\mathbf{w}}_t)$  是指标函数。 $c(\hat{\mathbf{w}}_t)$  代表识别出来的物体相应的置信度得分，有的物体置信度不高所以权重就相应降低些。

在本实施例的一些可选的实现方式中，通过图像三元组与相应生成的句子之间的语义相关性优化所述字幕器，包括：提取预置的第三样本集，其中，每个第三样本包括查询图像、正图像和负图像，正图像与查询图像共享至少两个对象，而负图像和查询图像没有任何共同的对象；基于机器学习方法，从所述第三样本集中选取第三样本，以及执行以下第三训练步骤：将选取的第三样本中的查询图像、正图像和负图像分别输入所述字幕器，输出查询句子、正句子和负句子；计算查询句子和正句子的第一语义相似度以及计算查询句子和负句子的第二语义相似度；根据所述第一语义相似度和所述第二语义相似度计算自监督三重态损失；若所述自监督三重态损失小于预定损失阈值，则确定出所述字幕器训练完成。

若所述自监督三重态损失不小于预定损失阈值，则调整所述字幕器的相关参数使得所述自监督三重态损失减小，以及从所述第三样本集中重新选取第三样本，继续执行所述第三训练步骤。

在本实施例的一些可选的实现方式中，所述计算查询句子和正句子的第一语义相似度以及计算查询句子和负句子的第二语义相似度，包括：对于查询句子、正句子和负句子，分别计算句子中每一个词的基于对象的概率分布，进行最大池化操作，分别得到查询句子特征、正句子特征和负句子特征；计算查询句子特征和正句子特征的第一语义相似度以及计算查询句子特征和负句子特征的第二语义相似度。

自监督的三重态损失 (Self-supervised Triplet Loss)，在具有对象包含奖励的优化中，无论图像之间相似或不相似的关系如何，都独立地利用每个图像与相应生成的句子之间的语义相关性。从探索相对关系的想法出发，本公

开设计了一种自监督的三元组损失，以三元组的方式在语义上约束字幕器的学习，旨在保留句子之间的相对语义顺序。基于图像中识别的视觉对象构造每个图像三元组（由查询图像，正图像和负图像组成）。正图像与查询图像共享至少两个对象，而负图像和查询图像没有任何共同的对象。给定这样的图像三元组，优化字幕器以使查询图像的生成语句比负图像的生成语句更类似于正图像的生成语句。具体来说，假设有一组三元组集 $\mathcal{T}$ ，其中每个三元组 $\{I_i, I_i^+, I_i^-\}$ 包含一个查询图像 $I_i$ ，一个正图像 $I_i^+$ 和一个负图像 $I_i^-$ ，这样的三元组输入字幕器生成对应的句子三元组 $\{S_i, S_i^+, S_i^-\}$ ，旨在使得字幕器生成的句子三元组中 $S_i$ 在语义上更趋近 $S_i^+$ 而疏远 $S_i^-$ 。因此，自监督的三重态损失计算公式如下：

$$\mathcal{L}_{tri} = \max(0, \alpha - \|\mathcal{F}(\hat{S}_i) - \mathcal{F}(\hat{S}_i^-)\|_2^2 + \|\mathcal{F}(\hat{S}_i) - \mathcal{F}(\hat{S}_i^+)\|_2^2)$$

其中 $\alpha$ 表示边距， $\mathcal{F}(\hat{S}_i)$ 表示 $\hat{S}_i$ 的基于对象的句子特征，具体来说，在解码阶段，通过仅保留 1600 个对象的概率，预测的单词分布被进一步转换为基于对象的分布。接下来，沿着 $\hat{S}_i$ 的解码过程累积所有基于对象的分布，并对它们执行最大池化，以产生相应的基于对象的句子特征。

可选地，最终整个模型的训练可以在自批判序列训练综合了对抗性奖励（Adversarial Reward）、对象包含奖励（Object Inclusion Reward）和自监督三重态损失（Self-supervised Triplet Loss），总体的目标梯度公式近似为：

$$\nabla_{\theta} \mathcal{L}_R = -[\lambda_1 r_{adv}(w_{1:T}^s) + \lambda_2 r_{obj}(w_{1:T}^s) - b] \nabla_{\theta} \log p_{\theta}(w_{1:T}^s) + \lambda_3 \nabla_{\theta} \mathcal{L}_{tri}$$

其中 $w_{1:T}^s$ 表示抽样句子， $b$ 表示获得的对抗性和对象包含奖励的组合。 $\lambda_1$ 、 $\lambda_2$ 、 $\lambda_3$ 分别代表的是对抗性奖励、对象包含奖励和自监督三重态损失的权重，权重可以为 0。

继续参见图 3，图 3 是根据本实施例的用于生成字幕器的方法的应用场景的一个示意图。在图 3 的应用场景中，将查询图像、正图像、负图像输入字幕器的图像编码器 Faster R-CNN，得到对象集 $\{\text{tree}(\text{树})、\text{man}(\text{人})、\text{bench}(\text{长椅})、\text{grass}(\text{草})、\text{dog}(\text{狗})\dots\}$ 。将对象集根据预定对象集分组后，输入字幕器的句子解码器（图 3 中下面的两层 LSTM 结构），基于对象集的语义进行波束搜索解码，生成伪句子“a man sitting on a bench near a tree”等。将这些伪句子和对应的图像作为伪图像句子对集，用于训练字幕器（图 3 中上面的两层 LSTM 结构代表了句子解码器）。为了简便起见，可固定图像编码器的参数，仅训练句子解码器，也可在句子解码器训练完成后再训练图像编码

器，交替训练图像编码器和句子解码器，得到性能最佳的字幕器。训练过程中使用交叉熵的方式。得到的上面的两层 LSTM 结构的参数可以共享给下面的两层 LSTM 结构。

为了进一步优化字幕器，可引入对抗性奖励、对象包含奖励和自监督三重态损失。

1、对抗性奖励优化：将真句子 “a cow stands in the back of a large truck” 和字幕器生成的伪句子一起输入句子鉴别器进行判别。如果判别的准确率未达到 0.5，则按最小化对抗性损失的方向调整句子鉴别器的参数，然后再按最大化对抗性奖励的方向调整字幕器的参数。交替训练（调整）句子鉴别器和字幕器可以对字幕器进行优化。

2、对象包含奖励优化：计算识别出的对象在字幕器生成的伪句子中的包含程度。例如，识别出的对象包括 tree, man, bench。如果句子 1 包括 tree（置信度为 0.9），句子 2 包括 tree（置信度为 0.8）和 man（置信度为 0.7），则句子 2 的对象包含奖励要高于句子 1。训练的目的是尽量提高对象包含奖励，每次调整参数都能提高对象包含奖励。

3、自监督三重态损失优化：图 3 中输入的样本可以是三元组：查询图像、正图像和负图像。不同图像能生成不同的伪句子，通过比较查询句子、正句子和负句子之间的语义相似性，来确定自监督三重态损失。训练的目的是降低自监督三重态损失，使得正句子与查询句子语义更接近，负句子与查询句子的语义不相关。

在训练阶段，本公开采用自学习模式，通过交替进行伪字幕对生成和字幕器再训练这两个过程来优化整个模型，以达到循环迭代地改进字幕器的目的。

本公开提出了基于语义约束的自学习框架，深入研究了非配对图像字幕器的自学习思想。从建立一个伪句子生成和迭代优化的角度研究了这个问题，逐步提高了句子生成的质量。此外，将语义约束很好地集成到模型中，充分利用图像中物体的语义指导字幕器的训练，得到了先进的无监督字幕技术。

进一步参考图 4，其示出了用于输出字幕的方法的又一个实施例的流程 400。该用于输出字幕的方法的流程 400，包括以下步骤：

步骤 401，获取待处理的图像。

在本实施例中，用于输出字幕的方法运行于其上的电子设备（例如图 1

所示的服务器)可以通过有线连接方式或者无线连接方式从用户利用其进行字幕编辑的终端接收待处理的图像。待处理的图像可以是单独的图像,也可以是视频文件,服务器将视频分帧后得到待处理的图像。

步骤 402, 将图像输入字幕器中, 输出图像对应的字幕。

在本实施例中, 字幕器为根据步骤 201-205 的方法训练得到的。通过字幕器能够为图像自动配上字幕。可将字幕直接输出到图像上, 也可生成单独的文件, 返回给终端, 由终端根据用户的需求设置字幕的格式, 再输出到图像上。字幕器不仅可以输入字幕, 还可输出图像编码器识别出的对象, 可用于训练过程中的语义约束。

步骤 401-402 可与步骤 201-205 交替执行。步骤 401-402 生成的字幕可作为步骤 201-205 的训练样本。

从图 4 中可以看出, 与图 2 对应的实施例相比, 本实施例中的用于输出字幕的方法的流程 400 体现了字幕器的应用步骤。由此, 本实施例描述的方案可以通过字幕器生成训练样本, 再用于字幕器的训练, 交替地生成字幕和字幕器再训练, 可以优化字幕器, 提高生成字幕的准确性。

进一步参考图 5, 作为对上述各图所示方法的实现, 本公开提供了一种用于生成字幕器的装置的一个实施例, 该装置实施例与图 2 所示的方法实施例相对应, 该装置具体可以应用于各种电子设备中。

如图 5 所示, 本实施例的用于生成字幕器的装置 500 包括: 获取单元 501、编码单元 502、分组单元 503、解码单元 504 和训练单元 505。其中, 获取单元 501, 被配置成获取样本图像集; 编码单元 502, 被配置成将所述样本图像集输入句子生成器的图像编码器, 输出对象集; 分组单元 503, 被配置成将所述对象集分组成第一对象集和第二对象集, 其中, 所述第一对象集为被包含在预定对象集内的对象集, 所述第二对象集为被排除在预定对象集外的对象集; 解码单元 504, 被配置成将所述图像编码器输出的对象集输入句子生成器的句子解码器, 在解码步骤中以所述第一对象集、所述第二对象集为约束条件进行波束搜索, 生成伪图像句子对集; 训练单元 505, 被配置成将所述伪图像句子对集作为样本集训练所述句子生成器, 得到字幕器。

在本实施例中, 用于生成字幕器的装置 500 的获取单元 501、编码单元 502、分组单元 503、解码单元 504 和训练单元 505 的具体处理可以参考图 2 对应实施例中的步骤 201、步骤 202、步骤 203、步骤 204 和步骤 205。

在本实施例的一些可选的实现方式中，该装置还包括优化单元（附图中未示出），被配置成：通过以下至少一种方式优化字幕器：通过句子鉴别器对字幕器进行对抗式训练来优化字幕器；通过字幕器识别出的对象在字幕器输出的句子中的包含程度优化字幕器；通过图像三元组与相应生成的句子之间的语义相关性优化字幕器，其中，图像三元组包括查询图像，正图像和负图像。

在本实施例的一些可选的实现方式中，优化单元进一步被配置成：提取预置的第一样本集，其中，每个第一样本包括图像和对应的真句子；提取预先建立的生成式对抗网络，其中，生成式对抗网络包括字幕器和句子鉴别器，字幕器用于对所输入的图像进行图像编码后再进行句子解码，得到伪句子，句子鉴别器用于确定所输入的句子是否为字幕器所输出的伪句子；基于机器学习装置，从第一样本集中选取第一样本，以及执行以下第一训练步骤：将选取的第一样本中的图像输入字幕器，输出伪句子；将伪句子和选取的第一样本中的真句子输入句子鉴别器，输入鉴别结果；根据输出的鉴别结果统计句子鉴别器的准确率；若准确率达到预设数值，则确定出字幕器训练完成。

在本实施例的一些可选的实现方式中，优化单元进一步被配置成：若准确率未达到预设数值，则计算句子鉴别器的对抗性损失，调整句子鉴别器的相关参数使得对抗性损失减小，以及从第一样本集中重新选取第一样本，继续执行第一训练步骤。

在本实施例的一些可选的实现方式中，优化单元进一步被配置成：若准确率未达到预设数值，则计算字幕器的对抗性奖励，调整字幕器的相关参数使得对抗性奖励增大，以及从第一样本集中重新选取第一样本，继续执行第一训练步骤。

在本实施例的一些可选的实现方式中，优化单元进一步被配置成：提取预置的第二样本集，其中，每个第二样本包括图像；基于机器学习装置，从第二样本集中选取样本，以及执行以下第二训练步骤：将选取的第二样本中的图像输入字幕器的图像编码器，输出样本对象集；将样本对象集输入字幕器的句子解码器，输出伪句子；计算伪句子中包含样本对象集中的样本对象的置信度均值分数，作为伪句子的对象包含奖励；若对象包含奖励达到预设包含奖励阈值，则确定出字幕器训练完成。

在本实施例的一些可选的实现方式中，优化单元进一步被配置成：若对

象包含奖励未达到预设包含奖励阈值，则调整字幕器的相关参数使得对象包含奖励增大，以及从第二样本集中重新选取第二样本，继续执行第二训练步骤。

在本实施例的一些可选的实现方式中，优化单元进一步被配置成：提取预置的第三样本集，其中，每个第三样本包括查询图像、正图像和负图像，正图像与查询图像共享至少两个对象，而负图像和查询图像没有任何共同的对象；基于机器学习装置，从第三样本集中选取第三样本，以及执行以下第三训练步骤：将选取的第三样本中的查询图像、正图像和负图像分别输入字幕器，输出查询句子、正句子和负句子；计算查询句子和正句子的第一语义相似度以及计算查询句子和负句子的第二语义相似度；根据第一语义相似度和第二语义相似度计算自监督三重态损失；若自监督三重态损失小于预定损失阈值，则确定出字幕器训练完成。

在本实施例的一些可选的实现方式中，优化单元进一步被配置成：若自监督三重态损失不小于预定损失阈值，则调整字幕器的相关参数使得自监督三重态损失减小，以及从第三样本集中重新选取第三样本，继续执行第三训练步骤。

在本实施例的一些可选的实现方式中，优化单元进一步被配置成：对于查询句子、正句子和负句子，分别计算句子中每一个词的基于对象的概率分布，进行最大池化操作，分别得到查询句子特征、正句子特征和负句子特征；计算查询句子特征和正句子特征的第一语义相似度以及计算查询句子特征和负句子特征的第二语义相似度。

在本实施例的一些可选的实现方式中，优化单元进一步被配置成：若对抗性奖励、对象包含奖励和自监督三重态损失的加权和大于预定目标值，则调整字幕器的相关参数使得加权和减小。

在本实施例的一些可选的实现方式中，图像编码器包括具有区域级别注意机制的两层 LSTM，其中，第一层 LSTM 充当自上而下的注意模块，根据上下文信息计算对象级别的注意，而第二层 LSTM 是用于生成句子的语言模型。

进一步参考图 6，作为对上述各图所示方法的实现，本公开提供了一种用于输出字幕的装置的一个实施例，该装置实施例与图 4 所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

如图 6 所示,本实施例的用于输出字幕的装置 600 包括:获取单元 601,被配置成获取待处理的图像;输出单元 602,被配置成将图像输入采用如装置 500 生成的字幕器中,输出图像对应的字幕。

根据本公开的实施例,本公开还提供了一种电子设备、一种可读存储介质。

图 7 示出了可以用来实施本公开的实施例的示例电子设备 700 的示意性框图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或要求的本公开的实现。

如图 7 所示,设备 700 包括计算单元 701,其可以根据存储在只读存储器 (ROM) 702 中的计算机程序或者从存储单元 708 加载到随机访问存储器 (RAM) 703 中的计算机程序,来执行各种适当的动作和处理。在 RAM 703 中,还可存储设备 700 操作所需的各种程序和数据。计算单元 701、ROM 702 以及 RAM 703 通过总线 704 彼此相连。输入/输出 (I/O) 接口 705 也连接至总线 704。

设备 700 中的多个部件连接至 I/O 接口 705,包括:输入单元 706,例如键盘、鼠标等;输出单元 707,例如各种类型的显示器、扬声器等;存储单元 708,例如磁盘、光盘等;以及通信单元 709,例如网卡、调制解调器、无线通信收发机等。通信单元 709 允许设备 700 通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

计算单元 701 可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元 701 的一些示例包括但不限于中央处理单元 (CPU)、图形处理单元 (GPU)、各种专用的人工智能 (AI) 计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器 (DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元 701 执行上文所描述的各个方法和处理,例如方法用于生成字幕器。例如,在一些实施例中,方法用于生成字幕器可被实现为计算机软件程序,其被有形地包含于机器可读介质,例如存储单元 708。在一些实施例中,计算机程序的部分或者全部可以经由 ROM 702 和/或通信单元

709 而被载入和/或安装到设备 700 上。当计算机程序加载到 RAM 703 并由计算单元 701 执行时，可以执行上文描述的方法用于生成字幕器的一个或多个步骤。备选地，在其他实施例中，计算单元 701 可以通过其他任何适当的方式（例如，借助于固件）而被配置为执行方法用于生成字幕器。

本公开的实施例提供的用于生成字幕器的方法和装置以及用于输出字幕的方法和装置，旨在为图像字幕提供无监督的解决方案。与现有的图像字幕方法严重依赖大量的图像句子对进行训练不同，本公开通过在自学方式中学习图像字幕器来消除这种依赖性。可以用非配对的图像和句子数据训练字幕器，追求更多真实的场景。

本文中以上描述的系统和技术各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列（FPGA）、专用集成电路（ASIC）、专用标准产品（ASSP）、芯片上系统的系统（SOC）、负载可编程逻辑设备（CPLD）、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器或控制器，使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦除可编程

只读存储器 (EPROM 或快闪存储器)、光纤、便捷式紧凑盘只读存储器 (CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：用于向用户显示信息的显示装置 (例如，CRT (阴极射线管) 或者 LCD (液晶显示器) 监视器)；以及键盘和指向装置 (例如，鼠标或者轨迹球)，用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈 (例如，视觉反馈、听觉反馈、或者触觉反馈)；并且可以用任何形式 (包括声输入、语音输入、或者触觉输入) 来接收来自用户的输入。

可以将此处描述的系统和技术实施在包括后台部件的计算系统 (例如，作为数据服务器)、或者包括中间件部件的计算系统 (例如，应用服务器)、或者包括前端部件的计算系统 (例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信 (例如，通信网络) 来将系统的部件相互连接。通信网络的示例包括：局域网 (LAN)、广域网 (WAN) 和互联网。

计算机系统可以包括客户端和服务端。客户端和服务端一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务端关系的计算机程序来产生客户端和服务端的关系。服务端可以为分布式系统的服务端，或者是结合了区块链的服务端。服务端也可以是云服务器，或者是带人工智能技术的智能云计算服务器或智能云主机。服务端可以为分布式系统的服务端，或者是结合了区块链的服务端。服务端也可以是云服务器，或者是带人工智能技术的智能云计算服务器或智能云主机。

应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本公开中记载的各步骤可以并行地执行也可以顺序地执行也可以以不同的次序执行，只要能够实现本公开公开的技术方案所期望的结果，本文在此不进行限制。

上述具体实施方式，并不构成对本公开保护范围的限制。本领域技术人员应该明白的是，根据设计要求和因素，可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进

等，均应包含在本公开保护范围之内。

## 权利要求书

1、一种用于生成字幕器的方法，包括：

获取样本图像集；

将所述样本图像集输入句子生成器的图像编码器，输出对象集；

将所述对象集分组成第一对象集和第二对象集，其中，所述第一对象集为被包含在预定对象集内的对象集，所述第二对象集为被排除在预定对象集外的对象集；

将所述图像编码器输出的对象集输入句子生成器的句子解码器，在解码步骤中以所述第一对象集、所述第二对象集为约束条件进行波束搜索，生成伪图像句子对集；

将所述伪图像句子对集作为样本集训练所述句子生成器，得到字幕器。

2、根据权利要求1所述的方法，其中，所述方法还包括：

通过以下至少一种方式优化所述字幕器：

通过句子鉴别器对所述字幕器进行对抗式训练来优化所述字幕器；

通过所述字幕器识别出的对象在所述字幕器输出的句子中的包含程度优化所述字幕器；

通过图像三元组与相应生成的句子之间的语义相关性优化所述字幕器，其中，图像三元组包括查询图像，正图像和负图像。

3、根据权利要求2所述的方法，其中，所述通过句子鉴别器对所述字幕器进行对抗式训练来优化所述字幕器，包括：

提取预置的第一样本集，其中，每个第一样本包括图像和对应的真句子；

提取预先建立的生成式对抗网络，其中，所述生成式对抗网络包括字幕器和句子鉴别器，所述字幕器用于对所输入的图像进行图像编码后再进行句子解码，得到伪句子，所述句子鉴别器用于确定所输入的句子是否为所述字幕器所输出的伪句子；

基于机器学习方法，从所述第一样本集中选取第一样本，以及执行以下第一训练步骤：将选取的第一样本中的图像输入所述字幕器，输出伪句子；将所述伪句子和选取的第一样本中的真句子输入所述句子鉴别器，输入鉴别结果；根据输出的鉴别结果统计所述句子鉴别器的准确率；若所述准确率达

到预设数值，则确定出所述字幕器训练完成。

4、根据权利要求3所述的方法，其中，所述方法还包括：

若所述准确率未达到所述预设数值，则计算所述句子鉴别器的对抗性损失，调整所述句子鉴别器的相关参数使得所述对抗性损失减小，以及从所述第一样本集中重新选取第一样本，继续执行所述第一训练步骤。

5、根据权利要求3所述的方法，其中，所述方法还包括：

若所述准确率未达到所述预设数值，则计算所述字幕器的对抗性奖励，调整所述字幕器的相关参数使得所述对抗性奖励增大，以及从所述第一样本集中重新选取第一样本，继续执行所述第一训练步骤。

6、根据权利要求2所述的方法，其中，所述通过所述字幕器识别出的对象在所述字幕器输出的句子中的包含程度优化所述字幕器，包括：

提取预置的第二样本集，其中，每个第二样本包括图像；

基于机器学习方法，从所述第二样本集中选取样本，以及执行以下第二训练步骤：将选取的第二样本中的图像输入所述字幕器的图像编码器，输出样本对象集；将所述样本对象集输入字幕器的句子解码器，输出伪句子；计算所述伪句子中包含所述样本对象集中的样本对象的置信度均值分数，作为所述伪句子的对象包含奖励；若所述对象包含奖励达到预设包含奖励阈值，则确定出所述字幕器训练完成。

7、根据权利要求6所述的方法，其中，所述方法还包括：

若所述对象包含奖励未达到所述预设包含奖励阈值，则调整所述字幕器的相关参数使得所述对象包含奖励增大，以及从所述第二样本集中重新选取第二样本，继续执行所述第二训练步骤。

8、根据权利要求2所述的方法，其中，所述通过图像三元组与相应生成的句子之间的语义相关性优化所述字幕器，包括：

提取预置的第三样本集，其中，每个第三样本包括查询图像、正图像和负图像，正图像与查询图像共享至少两个对象，而负图像和查询图像没有任

何共同的对象；

基于机器学习方法，从所述第三样本集中选取第三样本，以及执行以下第三训练步骤：将选取的第三样本中的查询图像、正图像和负图像分别输入所述字幕器，输出查询句子、正句子和负句子；计算查询句子和正句子的第一语义相似度以及计算查询句子和负句子的第二语义相似度；根据所述第一语义相似度和所述第二语义相似度计算自监督三重态损失；若所述自监督三重态损失小于预定损失阈值，则确定出所述字幕器训练完成。

9、根据权利要求 8 所述的方法，其中，所述方法还包括：

若所述自监督三重态损失不小于所述预定损失阈值，则调整所述字幕器的相关参数使得所述自监督三重态损失减小，以及从所述第三样本集中重新选取第三样本，继续执行所述第三训练步骤。

10、根据权利要求 8 所述的方法，其中，所述计算查询句子和正句子的第一语义相似度以及计算查询句子和负句子的第二语义相似度，包括：

对于查询句子、正句子和负句子，分别计算句子中每一个词的基于对象的概率分布，进行最大池化操作，分别得到查询句子特征、正句子特征和负句子特征；

计算查询句子特征和正句子特征的第一语义相似度以及计算查询句子特征和负句子特征的第二语义相似度。

11、根据权利要求 2-10 中任一项所述的方法，其中，所述方法还包括：

若对抗性奖励、对象包含奖励和自监督三重态损失的加权和大于预定目标值，则调整字幕器的相关参数使得所述加权和减小。

12、根据权利要求 1-10 中任一项所述的方法，其中，所述图像编码器包括具有区域级别注意机制的两层 LSTM，其中，第一层 LSTM 充当自上而下的注意模块，根据上下文信息计算对象级别的注意，而第二层 LSTM 是用于生成句子的语言模型。

13、一种用于输出字幕的方法，包括：

获取待处理的图像；

将所述图像输入采用如权利要求 1-12 中任一项所述的方法生成的字幕器中，输出所述图像对应的字幕。

14、一种用于生成字幕器的装置，包括：

获取单元，被配置成获取样本图像集；

编码单元，被配置成将所述样本图像集输入句子生成器的图像编码器，输出对象集；

分组单元，被配置成将所述对象集分组成第一对象集和第二对象集，其中，所述第一对象集为被包含在预定对象集内的对象集，所述第二对象集为被排除在预定对象集外的对象集；

解码单元，被配置成将所述图像编码器输出的对象集输入句子生成器的句子解码器，在解码步骤中以所述第一对象集、所述第二对象集为约束条件进行波束搜索，生成伪图像句子对集；

训练单元，被配置成将所述伪图像句子对集作为样本集训练所述句子生成器，得到字幕器。

15、一种用于输出字幕的装置，包括：

获取单元，被配置成获取待处理的图像；

输出单元，被配置成将所述图像输入采用如权利要求 1-12 中任一项所述的方法生成的字幕器中，输出所述图像对应的字幕。

16、一种电子设备，包括：

一个或多个处理器；

存储装置，其上存储有一个或多个计算机程序，

当所述一个或多个计算机程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求 1-13 中任一项所述的方法。

17、一种计算机可读介质，其上存储有计算机程序，其中，所述计算机程序被处理器执行时实现如权利要求 1-13 中任一项所述的方法。