(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 9 月 10 日 (10.09.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/177673 A1

(51) 国际专利分类号:
G06K 9/00 (2006.01) *G06K 9/62* (2006.01)

(21) 国际申请号: PCT/CN2020/077481

(22) 国际申请日: 2020 年 3 月 2 日 (02.03.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910165102.1 2019年3月5日 (05.03.2019) CN

(71) 申请人: 腾 讯 科 技 (深 圳) 有 限 公 司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(72) 发明人: 陈振方 (CHEN, Zhenfang); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 马林(MA, Lin); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 罗文寒 (LUO, Wenhan); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 刘威 (LIU, Wei); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(74) 代理人: 北京三高永信知识产权代理有限责任公司(BEIJING SAN GAO YONG XIN INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市海淀区学院路蓟门里和景园A座1单元102室, Beijing 100088 (CN)。

WO 2020/177673 A1

(54) **Title:** VIDEO SEQUENCE SELECTION METHOD, COMPUTER DEVICE AND STORAGE MEDIUM

(54) 发明名称: 一种视频序列选择的方法、计算机设备及存储介质

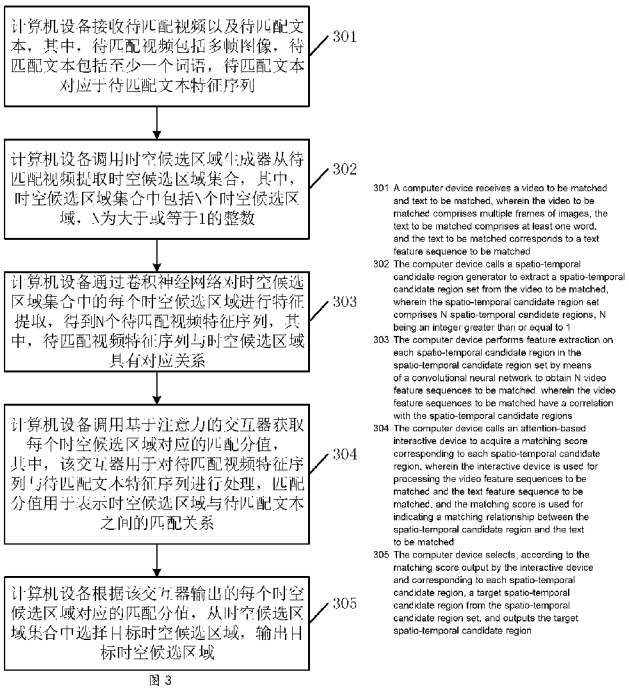

图 3

(57) **Abstract:** Disclosed is a video sequence selection method, applied to a computer device. The method comprises: receiving a video to be matched and text to be matched, wherein the text to be matched corresponds to a text feature sequence to be matched; calling a spatio-temporal candidate region generator to extract a spatio-temporal candidate region set from the video to be matched, wherein the spatio-temporal candidate region set comprises N spatio-temporal candidate regions; performing feature extraction on each spatio-temporal candidate region by means of a convolutional neural network to obtain N video feature sequences to be matched; calling

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

an attention-based interactive device to acquire a matching score corresponding to each spatio-temporal candidate region, wherein the matching score is used for indicating a matching relationship between the spatio-temporal candidate region and the text to be matched; and selecting, according to the matching score corresponding to each spatio-temporal candidate region, a target spatio-temporal candidate region from the spatio-temporal candidate region set, and outputting the target spatio-temporal candidate region. In the present application, a time sequence correlation between a video and text is taken into consideration during matching, thereby improving the matching degree of a video sequence and the text.

(57) 摘要：本申请公开了一种视频序列选择的方法，应用于计算机设备，包括：接收待匹配视频以及待匹配文本，待匹配文本对应于待匹配文本特征序列；调用时空候选区域生成器从待匹配视频中提取时空候选区域集合，时空候选区域集合中包括N个时空候选区域；通过卷积神经网络对每个时空候选区域进行特征提取，得到N个待匹配视频特征序列；调用基于注意力的交互器获取每个时空候选区域对应的匹配分值，匹配分值用于表示时空候选区域与待匹配文本之间的匹配关系；根据每个时空候选区域对应的匹配分值，从时空候选区域集合中选择目标时空候选区域，输出目标时空候选区域。本申请在匹配的时候考虑到了视频与文本在时序上的关联性，从而提升视频序列与文本的匹配度。

一种视频序列选择的方法、计算机设备及存储介质

本申请要求于2019年03月05日提交的申请号为201910165102.1、发明名称为“一种视频序列选择的方法、模型训练的方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及人工智能技术领域，尤其涉及一种视频序列选择的方法、计算机设备及存储介质。

背景技术

人工智能属于计算机科学的一个分支，主要目标是使机器能够胜任一些通常需要人类智能才能完成的复杂工作。在人工智能得到愈加广泛重视的背景下，随着计算机网络技术、多媒体技术以及数字传输技术的不断发展，以及摄像机、手机以及平板电脑等数码设备的不断普及，视频的数据量急剧增长。面对海量的视频，如何有效地对其进行处理，从而使用户能够迅速获取想要的信息，是当前研究和应用的关键问题。

目前，在提取视频中所需的内容时，通常是对视频中单帧图像和文本分别进行编码，然后将文本与每帧图像进行匹配，从而得到每帧图像与文本的匹配结果，再根据匹配结果得到单帧图像在视频中的空间位置，最后将这些空间位置串联起来得到一个与文本关联的视频序列。

然而，采用上述方式生成的视频序列虽然与文本之间具有关联性，但是仅仅考虑到单帧图像与文本之间的匹配关系，导致输出的视频序列与文本的匹配度较低，不利于对视频内容的理解。

发明内容

本申请实施例提供了一种视频序列选择的方法、、计算机设备及存储介质，由于时空候选区域包括了图像在时间和空间上的关系，因此，在匹配的时候考虑到了视频与文本在时序上的关联性，即考虑了视频时序信息对视频序列以及文本的影响，从而提升了输出的视频序列与文本的匹配度，进而有利于更好地理解视频内容。

有鉴于此，本申请第一方面提供一种视频序列选择的方法，所述方法应用于计算机设备，包括：

所述计算机设备接收待匹配视频以及待匹配文本，其中，所述待匹配视频包括多帧图像，所述待匹配文本包括至少一个词语，所述待匹配文本对应于待匹配文本特征序列；

所述计算机设备调用时空候选区域生成器从所述待匹配视频中提取时空候选区域集合，其中，所述时空候选区域集合中包括N个时空候选区域，所述N

为大于或等于 1 的整数，一个时空候选区域为一个视频序列；

所述计算机设备通过卷积神经网络对所述时空候选区域集合中的每个时空候选区域进行特征提取，得到 N 个待匹配视频特征序列，其中，所述待匹配视频特征序列与所述时空候选区域具有对应关系；

所述计算机设备调用基于注意力的交互器获取所述每个时空候选区域对应的匹配分值，其中，所述交互器用于对所述待匹配视频特征序列与所述待匹配文本特征序列进行处理，所述匹配分值用于表示所述时空候选区域与所述待匹配文本之间的匹配关系；

所述计算机设备根据所述交互器输出的所述每个时空候选区域对应的匹配分值，从所述时空候选区域集合中选择目标时空候选区域，输出所述目标时空候选区域。

本申请第二方面提供一种视频序列选择装置，包括：

获取模块，用于接收待匹配视频以及待匹配文本，其中，所述待匹配视频包括多帧图像，所述待匹配文本包括至少一个词语，所述待匹配文本对应于待匹配文本特征序列；

生成模块，用于调用时空候选区域生成器从所述待匹配视频中提取时空候选区域集合，其中，所述时空候选区域集合中包括 N 个时空候选区域，所述 N 为大于或等于 1 的整数，一个时空候选区域为一个视频序列；

编码模块，用于通过卷积神经网络对所述时空候选区域集合中的每个时空候选区域进行特征提取，得到 N 个待匹配视频特征序列，其中，所述待匹配视频特征序列与所述时空候选区域具有对应关系；

所述获取模块，还用于调用基于注意力的交互器获取所述每个时空候选区域对应的匹配分值，其中，所述交互器用于对所述待匹配视频特征序列与所述待匹配文本特征序列进行处理，所述匹配分值用于表示所述时空候选区域与所述待匹配文本之间的匹配关系；

选择模块，用于根据所述交互器输出的所述每个时空候选区域对应的匹配分值，从所述时空候选区域集合中选择目标时空候选区域，输出所述目标时空候选区域。

在一种可能的设计中，在本申请实施例的第二方面的第一种实现方式中，

所述生成模块，用于调用所述时空候选区域生成器获取所述待匹配视频中每帧图像的候选区域以及置信度得分，其中，每个候选区域对应一个置信度得分；获取所述待匹配视频中相邻两帧图像之间的重合度；根据所述每帧图像的候选区域、所述置信度得分以及所述重合度，生成所述时空候选区域集合。

在一种可能的设计中，在本申请实施例的第二方面的第二种实现方式中，

所述获取模块，用于对于所述每个时空候选区域，调用所述交互器的编码器对所述时空候选区域对应的待匹配视频特征序列进行编码，得到视觉特征集合，其中，所述视觉特征集合包括至少一个视觉特征；调用所述交互器的编码器对所述待匹配文本特征序列进行编码，得到文本特征集合，其中，所述文本特征集合包括至少一个文本特征；根据所述视觉特征集合以及所述文本特征集

合，确定视觉文本特征集合，其中，所述视觉文本特征集合包括至少一个视觉文本特征，所述视觉文本特征表示基于视觉特征的文本特征；根据所述视觉文本特征集合以及所述视觉特征集合，确定所述时空候选区域对应的匹配分值。

在一种可能的设计中，在本申请实施例的第二方面的第三种实现方式中，

所述获取模块，用于采用如下方式计算所述视觉特征集合：

$$H_p=\left\{h_t^p\right\}_{t=1}^{t_p};$$

$$h_t^p=LSTM_p\left(f_t^p,h_{t-1}^p\right);$$

其中，所述H_p表示所述视觉特征集合，所述h_t^p表示所述视觉特征集合中的第 t 个视觉特征，所述t_p表示所述时空候选区域的时间步数，所述h_{t-1}^p表示所述视觉特征集合中的第（t-1）个视觉特征，所述$LSTM_p(\ )$表示第一长短期记忆网络 LSTM 编码器，所述f_t^p表示所述待匹配视频特征序列中的第 t 行特征；

采用如下方式计算所述文本特征集合：

$$H_q=\left\{h_t^q\right\}_{t=1}^{t_q};$$

$$h_t^q=LSTM_q\left(f_t^q,h_{t-1}^q\right);$$

其中，所述H_q表示所述文本特征集合，所述h_t^q表示所述文本特征集合中的第 t 个文本特征，所述t_q表示所述待匹配文本的词语数量，所述h_{t-1}^q表示所述文本特征集合中的第（t-1）个文本特征，所述$LSTM_q(\ )$表示第二 LSTM 编码器，所述f_t^q表示所述待匹配文本特征序列中的第 t 行特征。

在一种可能的设计中，在本申请实施例的第二方面的第四种实现方式中，

所述获取模块，用于调用所述交互器执行根据所述视觉特征集合以及所述文本特征集合，计算视觉特征对应文本特征的注意力权重；根据所述注意力权重，计算所述视觉特征对应所述文本特征的归一化注意力权重；根据所述归一化注意力权重以及所述文本特征，计算视觉文本特征集合。

在一种可能的设计中，在本申请实施例的第二方面的第五种实现方式中，

所述获取模块，用于采用如下方式计算所述注意力权重：

$$e_{i,j}=w^T\tanh\left(W^qh_j^q+W^ph_i^p+b_1\right)+b_2;$$

其中，所述$e_{i,j}$表示第 i 个视觉特征对应第 j 个文本特征的注意力权重，所述h_j^q表示所述第 j 个文本特征，所述h_i^p表示所述第 i 个视觉特征，所述w^T表示第一模型参数，所述W^q表示第二模型参数，所述W^p表示第三模型参数，所述b_1表示第四模型参数，所述b_2表示第五模型参数，所述$\tanh(\ )$表示双曲正切函数；

采用如下方式计算所述归一化注意力权重：

$$a_{i,j}=\frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{t_q}\exp\left(e_{i,k}\right)};$$

其中，所述$a_{i,j}$表示所述第 i 个视觉特征对应所述第 j 个文本特征的归一化注意力权重，所述t_q表示所述待匹配文本的词语数量，所述k表示所述待匹配文

本中的第k个词语，所述k为大于或等于1，且小于或等于所述t_q的整数，所述exp( )表示指数函数；

采用如下方式计算所述视觉文本特征集合：

$$H_{qp}=\left\{h_{qp}\right\}_{t=1}^{t_p};$$

$$h_{qp}=\sum_{j=1}^{t_q}a_{i,j}h_j^q;$$

其中，所述H_{qp}表示所述视觉文本特征集合，所述t_p表示所述时空候选区域的时间步数，所述h_{qp}表示视觉文本特征。

在一种可能的设计中，在本申请实施例的第二方面的第六种实现方式中，

所述获取模块，用于采用如下方式计算所述匹配分值：

$$s(q,p)=\frac{1}{t_p}\sum_{i=1}^{t_p}s_i\left(h_i^p,h_i^{qp}\right);$$

$$s_i\left(h_i^p,h_i^{qp}\right)=\phi\left(h_i^p,h_i^{qp}\right);$$

其中，所述$s(q,p)$表示所述时空候选区域对应的匹配分值，所述$s_i\left(h_i^p,h_i^{qp}\right)$表示第 i 个时间步数对应的视觉特征和视觉文本特征之间的匹配子分值，所述h_i^{qp}表示所述第 i 个时间步数对应的视觉文本特征，所述h_i^p表示所述第 i 个时间步数对应的视觉特征，所述$\phi(\ )$表示相似度计算函数。

本申请第三方面提供一种模型训练装置，包括：

获取模块，用于获取第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，其中，所述第一待训练视频与所述第一待训练文本具有匹配关系，且所述第一待训练视频与所述第二待训练文本不具有匹配关系，所述第二待训练视频与所述第二待训练文本具有匹配关系，且所述第二待训练视频与所述第一待训练文本不具有匹配关系；

确定模块，用于根据所述获取模块获取的所述第一待训练视频、所述第二待训练视频、所述第一待训练文本以及所述第二待训练文本，确定排列损失函数，其中，所述排列损失函数用于对所述第一待训练视频以及所述第二待训练文本进行处理，并对所述第二待训练视频以及所述第一待训练文本进行处理；

所述确定模块，还用于根据所述获取模块获取的所述第一待训练视频、所述第二待训练视频、所述第一待训练文本以及所述第二待训练文本，确定多样性损失函数，其中，所述多样性损失函数用于对所述第一待训练视频以及所述第一待训练文本进行处理，并对所述第二待训练视频以及所述第二待训练文本进行处理；

所述确定模块，还用于根据所述排列损失函数以及所述多样性损失函数，确定目标损失函数；

训练模块，用于采用所述确定模块确定的所述目标损失函数对待训练的交互器进行训练，得到基于注意力的交互器，其中，所述交互器用于输出待匹配

视频与待匹配文本的匹配分值。

在一种可能的设计中，在本申请实施例的第三方面的第一种实现方式中，

所述确定模块，用于获取所述第一待训练视频中的第一时空候选区域集合，以及获取所述第二待训练视频中的第二时空候选区域集合，其中，所述第一时空候选区域集合包括至少一个第一时空候选区域，所述第一时空候选区域为视频序列，所述第二时空候选区域集合包括至少一个第二时空候选区域，所述第二时空候选区域为视频序列；根据所述第一待训练文本以及所述第二时空候选区域集合，计算第一匹配分值；根据所述第二待训练文本以及所述第一时空候选区域集合，计算第二匹配分值；根据所述第一待训练文本以及所述第一时空候选区域集合，计算第三匹配分值；根据所述第一匹配分值、所述第二匹配分值以及所述第三匹配分值，确定所述排列损失函数。

在一种可能的设计中，在本申请实施例的第三方面的第二种实现方式中，

所述确定模块，用于根据第一时空候选区域集合以及所述第一待训练文本，确定匹配行为分布，其中，所述第一时空候选区域集合是根据所述第一待训练视频生成的，所述匹配行为分布表示所述第一时空候选区域集合中每个第一时空候选区域与所述第一待训练文本之间的匹配关系；对所述匹配行为分布进行归一化处理，得到目标匹配行为分布；根据所述目标匹配行为分布确定所述多样性损失函数。

在一种可能的设计中，在本申请实施例的第三方面的第三种实现方式中，

所述确定模块，用于获取控制系数，根据所述控制系数、所述排列损失函数以及所述多样性损失函数，确定所述目标损失函数。

本申请第四方面提供一种计算机设备，包括：存储器、收发器、处理器以及总线系统；

其中，所述存储器用于存储程序；

所述处理器用于执行所述存储器中的程序，包括如下步骤：

接收待匹配视频以及待匹配文本，其中，所述待匹配视频包括多帧图像，所述待匹配文本包括至少一个词语，所述待匹配文本对应于待匹配文本特征序列；

调用时空候选区域生成器从所述待匹配视频中提取时空候选区域集合，其中，所述时空候选区域集合中包括 N 个时空候选区域，所述 N 为大于或等于 1 的整数，一个时空候选区域为一个视频序列；

通过卷积神经网络对所述时空候选区域集合中的每个时空候选区域进行特征提取，得到 N 个待匹配视频特征序列，其中，所述待匹配视频特征序列与所述时空候选区域具有对应关系；

调用基于注意力的交互器获取所述每个时空候选区域对应的匹配分值，其中，所述交互器用于对所述待匹配视频特征序列与所述待匹配文本特征序列进行处理，所述匹配分值用于表示所述时空候选区域与所述待匹配文本之间的匹配关系；

根据所述交互器输出的所述每个时空候选区域对应的匹配分值，从所述时

空候选区域集合中选择目标时空候选区域，输出所述目标时空候选区域；

所述总线系统用于连接所述存储器以及所述处理器，以使所述存储器以及所述处理器进行通信。

本申请的第五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指令，当其在计算机设备上运行时，使得计算机设备执行上述第一方面所述的视频序列选择的方法。

从以上技术方案可以看出，本申请实施例具有以下优点：

本申请实施例提供了一种视频序列选择的方法，首先接收待匹配视频以及待匹配文本，其中，待匹配视频包括多帧图像，待匹配文本包括至少一个词语，待匹配文本对应于待匹配文本特征序列，然后从待匹配视频中提取时空候选区域集合，接下来需要对时空候选区域集合中的每个时空候选区域进行特征提取，得到N个待匹配视频特征序列，其中，待匹配视频特征序列与时空候选区域具有对应关系，然后可以调用基于注意力的交互器获取每个时空候选区域对应的匹配分值，最后根据每个时空候选区域对应的匹配分值，从时空候选区域集合中选择目标时空候选区域，其中，一个时空候选区域为一个视频序列。通过上述方式，对视频中的时空候选区域和文本进行匹配，而不再是对视频中的每帧图像与文本进行匹配，这样操作的好处是，由于时空候选区域包括了图像在时间和空间上的关系，因此，在匹配的时候考虑到了视频与文本在时序上的关联性，即考虑了视频时序信息对视频序列以及文本的影响，从而提升了输出的视频序列与文本的匹配度，进而有利于更好地理解视频内容。

附图说明

图1为本申请实施例中视频序列选择系统的一个架构示意图；

图2为本申请实施例中视频序列选择系统的一个框架与流程示意图；

图3为本申请实施例中视频序列选择的方法一个实施例示意图；

图4为本申请实施例中提取时空候选区域的一个实施例示意图；

图5为本申请实施例中基于注意力机制的一个交互器结构示意图；

图6为本申请实施例中模型训练的方法一个实施例示意图；

图7为本申请实施例中视频序列选择装置一个实施例示意图；

图8为本申请实施例中模型训练装置一个实施例示意图；

图9为本申请实施例中服务器一个结构示意图。

具体实施方式

人工智能(Artificial Intelligence, AI)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最佳结果的理论、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。

人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

计算机视觉技术(Computer Vision, CV)计算机视觉是一门研究如何使机器“看”的科学，更进一步的说，就是指用摄影机和电脑代替人眼对目标进行识别、跟踪和测量等机器视觉，并进一步做图形处理，使电脑处理成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科，计算机视觉研究相关的理论和技术，试图建立能够从图像或者多维数据中获取信息的人工智能系统。计算机视觉技术通常包括图像处理、图像识别、图像语义理解、图像检索、光学字符识别（Optical Character Recognition, OCR)、视频处理、视频语义理解、视频内容/行为识别、三维物体重建、三维技术、虚拟现实、增强现实、同步定位与地图构建等技术，还包括常见的人脸识别、指纹识别等生物特征识别技术。

机器学习(Machine Learning, ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

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。

本申请实施例提供的方案涉及人工智能的计算机视觉和机器学习等技术，通过如下实施例进行说明：

本申请实施例提供了一种视频序列选择的方法、计算机设备及存储介质，由于时空候选区域包括了图像在时间和空间上的关系，因此，在匹配的时候考虑到了视频与文本在时序上的关联性，即考虑了视频时序信息对视频序列以及文本的影响，从而提升了输出的视频序列与文本的匹配度，进而有利于更好地理解视频内容。

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等（如果存在）是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解的是，这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例，能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“对应于”以及他们的任何变形，意图在于覆

盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备，不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

应该理解的是，本申请可以应用于视频内容理解和定位的场景，包含但不仅限于视频分类的场景，在视频类网站上进行快速检索的场景，以及在视频中快速定位的场景。采用本申请所提供的视频序列选择的方法，能够衡量文本和视频之间的匹配关系，从而实现给定一个句子和一段视频，即可输出一段视频序列的目的。比如，需要在一段美国职业篮球联赛（National Basketball Association，NBA）视频中提取与球员库里相关的视频序列，以此制成一段视频集锦。首先，通过本申请所提供的方法，会先生成多个时空候选区域，这些时空候选区域也就是视频序列，然后根据文本“Curie's three-point shot（库里的三分球）”，从这些时空候选区域中选择与该文本匹配度最高的时空候选区域作为目标时空候选区域，可以将目标时空候选区域记为视频序列1。类似地，如果需要提取多个时空候选区域，则可以再输入不同的文本，比如“Harden drives the ball（哈登带球过人）”，然后从这些时空候选区域中选择与该文本匹配度最高的时空候选区域作为目标时空候选区域，可以将目标时空候选区域记为视频序列2。如果需要制成视频集锦，那么可以对视频1和视频2进行拼接，得到最终的视频。

为了便于理解，本申请提出了一种视频序列选择的方法，该方法应用于图1所示的视频序列选择系统，请参阅图1，图1为本申请实施例中视频序列选择系统的一个架构示意图，如图1所示，本申请中所提供的视频序列选择方法通常应用于计算机设备，该计算机设备可以是服务器100，也可以是客户端，本申请将以应用于服务器为例介绍。

在结合图1的基础上，请参阅图2，图2为本申请实施例中视频序列选择系统的一个框架与流程示意图，如图2所示，服务器100获取从客户端输入的视频，可以理解的是，该视频也可以是预先存储于服务器100中的数据，此处不做限定。接下来，服务器100通过时空候选区域生成器从视频中提取多个时空候选区域，如图2中的时空候选区域A、时空候选区域B和时空候选区域C。用户可通过客户端输入一段句子，比如“A brown squirrel is playing with a blue ball on the floor（一只棕色松鼠正在地板上玩蓝色的球）”，采用基于注意力的交互器将该句子分别与时空候选区域A、时空候选区域B和时空候选区域C进行交互。得到句子与时空候选区域A的匹配值为60，句子与时空候选区域B的匹配值为50，句子与时空候选区域C的匹配值为90，于是将时空候选区域C作为目标时空候选区域输出，其中，时空候选区域C表现为一段视频序列。

此外，基于注意力的交互器是通过损失函数优化得到的，该损失函数可以包括一个排列损失函数和一个多样性函数，此处不做限定。其中，上述基于注意力的交互器在后文中也被称之为视频文本交互模型。

需要说明的是，客户端部署于终端设备110上，其中，终端设备110包含但不仅限于平板电脑、笔记本电脑、掌上电脑、手机、语音交互设备及个人电脑（personal computer，PC），此处不做限定。其中，语音交互设备包含但不

仅限于智能音响以及智能家电。

结合上述介绍，下面将对本申请中视频序列选择的方法进行介绍，请参阅图 3，本申请中视频序列选择的方法的一个实施例包括：

301、计算机设备接收待匹配视频以及待匹配文本，其中，待匹配视频包括多帧图像，待匹配文本包括至少一个词语，待匹配文本对应于待匹配文本特征序列；

本实施例中，视频序列选择装置首先需要获取待匹配视频以及待匹配文本，其中，该视频序列选择装置既可以部署于图 1 中所示的服务器 100 上，也可以部署于图 1 中所示的计算能力较强的终端设备 110 上，此处不做限定。待匹配视频包括多帧图像，待匹配文本包括至少一个词语，在获取到待匹配文本之后，可以采用词向量模型对该待匹配文本进行处理，实现将语言数学化，从而得到待匹配文本特征序列。待匹配文本特征序列中包括了至少一个词向量，词向量即为词语的向量表示。假设待匹配文本为句子 q，对句子 q 进行编码，得到待匹配文本特征序列 F_q，且 $F_q \in R^{t_q \times d_q}$，其中，$t_q$ 表示句子 q 中的词语数量，d_q 表示词向量的特征长度。

可以理解的是，待匹配视频的视频格式包含但不仅限于运动图像专家组（motion picture experts group，MPEG）格式、音频视频交错（audio video interleaved，AVI）、格式、高级流格式（advanced streaming format，ASF）、微软媒体视频（Windows media video，WMV）格式、第三代合作伙伴项目计划文件格式（3rd generation partnership project file format，3GP）、多媒体容器文件格式（multimedia container file format，MKV）、流媒体格式（flash video）以及视频容器可变比特率文件格式（RealMedia variable bitrate file format，RMVB）等。

可以理解的是，待匹配文本的语言类型包含但不仅限于中文、英文、日文、法文、德文以及阿拉伯语等。

302、计算机设备调用时空候选区域生成器从待匹配视频提取时空候选区域集合，其中，时空候选区域集合中包括 N 个时空候选区域，N 为大于或等于 1 的整数；

本实施例中，视频序列选择装置对待匹配视频进行处理，得到一系列时空候选区域，这一系列的时空候选区域称为时空候选区域集合，时空候选区域集合包括 N 个时空候选区域，可以表示为 $P=\{p_i\}_{i=1}^{N}$，N 即表示时空候选区域集合中的时空候选区域总数，一个时空候选区域即为一个视频序列。

303、计算机设备通过卷积神经网络对时空候选区域集合中的每个时空候选区域进行特征提取，得到 N 个待匹配视频特征序列，其中，待匹配视频特征序列与时空候选区域具有对应关系；

本实施例中，视频序列选择装置分别对时空候选区域集合中的每个时空候选区域进行编，以一个时空候选区域为例，假设该时空候选区域表示为 p，使用卷积神经网络对其提取相应的序列特征，即得到待匹配视频特征序列 F_p，且 $F_p \in R^{t_p \times d_p}$，其中，$t_p$ 表示视频的时间步数，时间步数表示视频被压缩后的固定

长度，d_p表示视频的特征长度，比如可以是 6048 或 4096 等，此处不做限定。每个时空候选区域对应一个待匹配视频特征序列。

304、计算机设备调用基于注意力的交互器获取每个时空候选区域对应的匹配分值，其中，该交互器用于对待匹配视频特征序列与待匹配文本特征序列进行处理，匹配分值用于表示时空候选区域与待匹配文本之间的匹配关系；

本实施例中，视频序列选择装置将每个待匹配视频特征序列以及待匹配文本特征序列输入至视频文本交互模型，由该视频文本交互模型输出相应的匹配分值。比如，时空候选区域集合包括 3 个时空候选区域，每个时空候选区域分别对应一个待匹配视频特征序列，比如时空候选区域 A 对应待匹配视频特征序列 A，时空候选区域 B 对应待匹配视频特征序列 B，时空候选区域 C 对应待匹配视频特征序列 C。此时，将待匹配视频特征序列 A 与待匹配文本特征序列输入至视频文本交互模型，由视频文本交互模型输出匹配分值 A。将待匹配视频特征序列 B 与待匹配文本特征序列输入至视频文本交互模型，由视频文本交互模型输出匹配分值 B。将待匹配视频特征序列 C 与待匹配文本特征序列输入至视频文本交互模型，由视频文本交互模型输出匹配分值 C。通常情况下，匹配分值越高，表示匹配关系越强。

305、计算机设备根据该交互器输出的每个时空候选区域对应的匹配分值，从时空候选区域集合中选择目标时空候选区域，输出目标时空候选区域。

本实施例中，视频序列选择装置根据每个时空候选区域对应的匹配分值，从中选择匹配分值最高的时空候选区域作为目标时空候选区域，并输出该目标时空候选区域，其中，目标时空候选区域即为一个视频序列。

本申请实施例提供了一种视频序列选择的方法，首先接收待匹配视频以及待匹配文本，其中，待匹配视频包括多帧图像，待匹配文本包括至少一个词语，待匹配文本对应于待匹配文本特征序列，然后从待匹配视频中提取时空候选区域集合，接下来需要对时空候选区域集合中的每个时空候选区域进行特征提取，得到 N 个待匹配视频特征序列，其中，待匹配视频特征序列与时空候选区域具有对应关系，然后可以通过视频文本交互模型获取每个时空候选区域对应的匹配分值，最后根据每个时空候选区域对应的匹配分值，从时空候选区域集合中选择目标时空候选区域，其中，一个时空候选区域为一个视频序列。通过上述方式，将视频中的时空候选区域和文本进行匹配，而不再是将视频中的每帧图像与文本进行匹配，这样操作的好处是，由于时空候选区域包括了图像在时间和空间上的关系，因此，在匹配的时候考虑到了视频与文本在时序上的关联性，即考虑了视频时序信息对视频序列以及文本的影响，从而提升了输出的视频序列与文本的匹配度，进而有利于更好地理解视频内容。

可选地，在上述图 3 对应的实施例的基础上，本申请实施例提供的视频序列选择的方法还包括一个可选实施例，在该可选实施例中，上述步骤 302 计算机设备调用时空候选区域生成器从待匹配视频提取时空候选区域集合，可以包括：

所述计算机设备调用所述时空候选区域生成器获取待匹配视频中每帧图像

的候选区域以及置信度得分，其中，每个候选区域对应一个置信度得分；获取待匹配视频中相邻两帧图像之间的重合度；根据每帧图像的候选区域、置信度得分以及重合度，生成时空候选区域集合。

本实施例中，介绍了一种根据待匹配视频生成时空候选区域集合的方式，首先，视频序列选择装置采用单帧的候选区域生成方法检测待匹配视频中的每一帧，由此得到每帧图像中的候选区域以及置信度得分，为了便于理解，请参阅图 4，图 4 为本申请实施例中提取时空候选区域的一个实施例示意图，如图 4 所示，在一帧图像中可以提取至少一个候选区域，需要注意的是，一帧图像是没有时序信息的，多帧图像就具有时序信息。候选区域 S1 为一个西瓜的图像，候选区域 S2 为一个菠萝的图像，由此可见，不同的候选区域其对应的边框大小也不同。以西瓜对应的候选区域为例，假设有 10 帧图像的置信度得分相近，且重合度高，则这 10 帧图像构成“西瓜”对应的时空候选区域。

其中，时空候选区域由一系列的边框$\{b_i\}_{i=1}^{T}$构成，b_i表示待匹配视频中第 i 帧中的一个候选区域，T 表示待匹配视频的图像总帧数。

其次，本申请实施例还提供了一种确定时空候选区域集合的方式，首先可以获取待匹配视频中每帧图像的候选区域以及置信度得分，每个候选区域对应一个置信度得分，然后获取待匹配视频中相邻两帧之间的重合度，最后根据每帧图像的候选区域、置信度得分以及重合度，生成时空候选区域集合。通过上述方式，结合了视频在时间和空间上的变化，同时结合了视频中外观信号和运动信号生成时空候选区域，由此，可以提升时空候选区域生成的准确性。

可选地，在上述图 3 对应的实施例的基础上，本申请实施例提供的视频序列选择的方法还包括另一个可选实施例，在该可选实施例中，上述步骤 304 所述计算机设备调用基于注意力的交互器获取所述每个时空候选区域对应的匹配分值，可以包括：

对于每个时空候选区域，所述计算机设备调用所述交互器的编码器对该时空候选区域对应的待匹配视频特征序列进行编码，得到视觉特征集合，其中，视觉特征集合包括至少一个视觉特征；所述计算机设备调用所述交互器的编码器对待匹配文本特征序列进行编码，得到文本特征集合，其中，文本特征集合包括至少一个文本特征；所述计算机设备调用所述交互器执行根据视觉特征集合以及文本特征集合，确定视觉文本特征集合，其中，视觉文本特征集合包括至少一个视觉文本特征，视觉文本特征表示基于视觉特征的文本特征；根据视觉文本特征集合以及视觉特征集合，确定该时空候选区域对应的匹配分值。

本实施例中，介绍了一种通过视频文本交互模型获取时空候选区域对应的匹配分值的实现方式。为了便于理解，请参阅图 5，图 5 为本申请实施例中基于注意力机制的一个交互器结构示意图，如图 5 所示，为了便于说明，下面将以一个时空候选区域为例进行说明，首先获取该时空候选区域对应的待匹配视频特征序列F_p以及待匹配文本特征序列F_q，然后使用两个基于长短期记忆网络（Long Short-Term Memory，LSTM）的循环神经网络作为编码器。其中，LSTM

编码器属于视频文本交互模型的组成部分。

然后模型训练装置将待匹配视频特征序列F_p输入至一个LSTM编码器，由这个LSTM编码器对待匹配视频特征序列F_p进行编码，得到视觉特征集合H_p，其中，视觉特征集合H_p包括t_p个视觉特征h^p。将待匹配文本特征序列F_q输入至另一个LSTM编码器，由另一个LSTM编码器对待匹配文本特征序列F_q进行编码，得到文本特征集合H_q，其中，文本特征集合H_q包括t_q个文本特征h^q。

接下来，模型训练装置采用注意力机制对文本特征集合H_q进行有针对性的加权求和，从而得到视觉文本特征集合H_{qp}，其中，视觉文本特征集合H_{qp}包括t_p个视觉文本特征h^{qp}，视觉文本特征即为视觉导向的文本特征。最后，模型训练装置对视觉文本特征集合H_{qp}中的视觉文本特征h^{qp}以及视觉特征集合H_p中视觉特征h^p进行计算，得到t_p个分值，对这t_p个分值求和即可得到该时空候选区域对应的匹配分值。

可以理解的是，时空候选区域集合中每个时空候选区域的处理方式均如上所述，由此，得到时空候选区域集合中各个时空候选区域的匹配分值，此处不再赘述其他时空候选区域的处理方式。

其次，本申请实施例还提供了一种获取时空候选区域对应的匹配分值的方式，分别通过视频文本交互模型的编码器对待匹配视频特征序列进行编码，得到视觉特征集合，并对待匹配文本特征序列进行编码，得到文本特征集合，然后根据视觉特征集合以及文本特征集合，确定视觉文本特征集合，最后，根据视觉文本特征集合以及视觉特征集合，确定时空候选区域对应的匹配分值。通过上述方式，利用交互机制将视频和文本进行特征融合，且能够刻画视频中每一个时空候选区域与文本之间的匹配关系，由此，在时间和空间上都能够实现与文本的匹配，进而提高对视频内容理解能力。

可选地，在上述图 3 对应的另一个可选实施例的基础上，本申请实施例提供的视频序列选择的方法还包括再一个可选实施例，在该可选实施例中，调用所述交互器的编码器对待匹配视频特征序列进行编码，得到视觉特征集合，可以包括：

采用如下方式计算视觉特征集合：

$$H_p = \left\{ h_t^p \right\}_{t=1}^{t_p};$$

$$h_t^p = LSTM_p\left(f_t^p, h_{t-1}^p \right);$$

其中，H_p表示视觉特征集合，h_t^p表示视觉特征集合中的第 t 个视觉特征，t_p表示时空候选区域的时间步数，h_{t-1}^p表示视觉特征集合中的第（t-1）个视觉特征，$LSTM_p(\ )$表示第一LSTM编码器，f_t^p表示待匹配视频特征序列中的第 t 行特征；

通过视频文本交互模型的编码器对待匹配文本特征序列进行编码，得到文本特征集合，可以包括：

采用如下方式计算文本特征集合：

$$H_q = \left\{ h_t^q \right\}_{t=1}^{t_q};$$

$$h_t^q = LSTM_q\left(f_t^q, h_{t-1}^q \right);$$

其中，H_q表示文本特征集合，h_t^q表示文本特征集合中的第 t 个文本特征，t_q表示待匹配文本的词语数量，h_{t-1}^q表示文本特征集合中的第（t-1）个文本特征，$LSTM_q(\ )$表示第二 LSTM 编码器，f_t^q表示待匹配文本特征序列中的第 t 行特征。

本实施例中，将介绍一种生成视觉特征集合以及生成文本特征集合的实现方式。下面介绍生成视觉特征集合的方式，即采用 LSTM 编码器对视觉特征集合H_p中的第(t-1)个视觉特征h_{t-1}^p以及待匹配视频特征序列F_p中的第 t 行特征f_t^p输入至第一 LSTM 编码器，由第一 LSTM 编码器输出第 t 个视觉特征h_t^p。当输出t_p个视觉特征h_t^p之后，即可得到视觉特征集合H_p。

下面介绍生成文本特征集合的方式，即采用 LSTM 编码器对文本特征集合H_q中的第（t-1）个文本特征h_{t-1}^q以及待匹配文本特征序列F_q中的第 t 行特征f_t^q输入至第二 LSTM 编码器，由第二 LSTM 编码器输出第 t 个文本特征h_t^q。当输出t_q个文本特征h_t^q之后，即可得到文本特征集合H_q。

其中，LSTM 编码器增加了对过去状态的过滤，从而可以选择哪些状态对当前更有影响，而不是简单的选择最近的状态。LSTM 编码器结构包含了遗忘门、学习门、记忆门以及使用门，长期记忆进入遗忘门，忘记它认为没有用处的内容。短期记忆和事件在学习门里合并到一起，并移除掉不必要的信息，作为学到的新信息。还没遗忘的长期记忆和刚学到的新信息会在记忆门里合并到一起，这个门将这两者放到一起，由于它叫记忆门，所以它会输出更新后的长期记忆。最后，使用门会决定要从之前知道的信息以及刚学到的信息中挑出什么来使用，从而作出预测，所以它也接受长期记忆和新信息的输入，将它们合并到一起并决定要输出什么。

再次，本申请实施例还提供了一种利用编码器对待匹配视频特征序列进行编码，得到视觉特征集合的实现方式，以及利用编码器对待匹配文本特征序列进行编码，得到文本特征集合的实现方式。通过上述方式，实现对特征序列的编码处理，并且采用 LSTM 编码器进行编码，由此可以处理和预测时间序列中间隔和延迟相对较长的重要事件，从而提升方案的可行性和可操作性。

可选地，在上述图 3 对应的另一个可选实施例的基础上，本申请实施例提供的视频序列选择的方法还包括又一个可选实施例，在该可选实施例中，计算机设备调用所述交互器执行根据视觉特征集合以及文本特征集合，确定视觉文本特征集合，可以包括：

所述计算机设备调用所述交互器执行根据视觉特征集合以及文本特征集合，计算视觉特征对应文本特征的注意力权重；根据该注意力权重，计算视觉特征对应文本特征的归一化注意力权重；根据归一化注意力权重以及文本特征，计算视觉文本特征集合。

本实施例中，介绍了一种生成视觉文本特征集合的方式。

可以理解的是，注意力机制（attention mechanism）是解决信息超载问题的主要手段，属于一种资源分配方案，将计算资源分配给更重要的任务，本申请采取的注意力机制可以是多头注意力、硬性注意力、键值对注意力或者结构化注意力。

多头注意力是利用多个查询，来平行地计算从输入信息中选取多个信息。每个注意力关注输入信息的不同部分。硬注意力，即基于注意力分布的所有输入信息的期望。还有一种注意力是只关注到一个位置上，叫做硬性注意力。硬性注意力有两种实现方式，一种是选取最高概率的输入信息。另一种硬性注意力可以通过在注意力分布式上随机采样的方式实现。键值对注意力可以用键值对格式来表示输入信息，其中“键”用来计算注意力分布，“值”用来生成选择的信息。结构化注意力要从输入信息中选取出和任务相关的信息，主动注意力是在所有输入信息上的多项分布，是一种扁平结构。如果输入信息本身具有层次结构，比如文本可以分为词、句子、段落以及篇章等不同粒度的层次，我们可以使用层次化的注意力来进行更好的信息选择。

再次，本申请实施例还提供了一种确定视觉文本特征集合的方式，即首先根据视觉特征集合以及文本特征集合，获取视觉特征对应文本特征的注意力权重，然后根据该注意力权重，获取视觉特征对应文本特征的归一化注意力权重，最后根据该归一化注意力权重以及文本特征，获取视觉文本特征集合。通过上述方式，充分利用注意力机制生成视觉文本特征，以此获取更多需要关注目标的视觉信息，而抑制其他无用信息，从而极大地提高了视觉信息处理的效率与准确性，能够从众多信息中选择出对当前任务目标更关键的信息。

可选地，在上述又一个可选实施例的基础上，本申请实施例提供的视频序列选择的方法还包括另一个可选实施例，在该可选实施例中，计算机设备调用所述交互器执行根据视觉特征集合以及文本特征集合，计算视觉特征对应文本特征的注意力权重，可以包括：

采用如下方式获取注意力权重：

$$e_{i,j} = w^T \tanh\left(W^q h_j^q + W^p h_i^p + b_1\right) + b_2 ;$$

其中，$e_{i,j}$ 表示第 i 个视觉特征对应第 j 个文本特征的注意力权重，h_j^q 表示第 j 个文本特征，h_i^p 表示第 i 个视觉特征，w^T 表示第一模型参数，W^q 表示第二模型参数，W^p 表示第三模型参数，b_1 表示第四模型参数，b_2 表示第五模型参数，$\tanh(\ )$ 表示双曲正切函数；

根据视觉特征对应文本特征的注意力权重，计算视觉特征对应文本特征的归一化注意力权重，可以包括：

采用如下方式计算归一化注意力权重：

$$a_{i,j} = \frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{t_q} \exp\left(e_{i,k}\right)} ;$$

其中，$a_{i,j}$ 表示第 i 个视觉特征对应第 j 个文本特征的归一化注意力权重，

t_q表示待匹配文本的词语数量，k表示待匹配文本中的第k个词语，k为大于或等于1，且小于或等于t_q的整数，exp( )表示指数函数；

根据该归一化注意力权重以及文本特征，计算视觉文本特征集合，包括：

采用如下方式计算视觉文本特征集合：

$$H_{qp}=\left\{h_{qp}\right\}_{t=1}^{t_p};$$

$$h_{qp}=\sum_{j=1}^{t_q}a_{i,j}h_j^q;$$

其中，H_{qp}表示视觉文本特征集合，t_p表示时空候选区域的时间步数，h_{qp}表示视觉文本特征。

本实施例中，介绍了一种计算视觉文本特征集合的实现方式。首先，从文本特征集合H_q中获取第 j 个文本特征h_j^q，并且从视觉特征集合H_p中获取第 i 个视觉特征h_i^p，采用如下方式计算注意力权重：

$$e_{i,j}=w^T\tanh\left(W^qh_j^q+W^ph_i^p+b_1\right)+b_2;$$

接下来，对注意力权重进行归一化处理，即采用如下方式计算归一化注意力权重：

$$a_{i,j}=\frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{t_q}\exp\left(e_{i,k}\right)};$$

最后采用如下方式计算视觉文本特征集合：

$$H_{qp}=\left\{h_{qp}\right\}_{t=1}^{t_p};$$

$$h_{qp}=\sum_{j=1}^{t_q}a_{i,j}h_j^q。$$

进一步地，本申请实施例还提供了一种计算视觉特征对应文本特征的注意力权重的实现方式，同时，提供了一种计算视觉特征对应文本特征的归一化注意力权重的实现方式，以及提供了一种计算视觉文本特征集合的实现方式。通过上述方式，为方案的实现提供了具体且可行的方式，由此，提升方案的实用性和可行性。

可选地，在上述任一个可选实施例的基础上，本申请实施例提供的视频序列选择的方法还包括另一个可选实施例，在该可选实施例中，计算机设备调用所述交互器执行根据视觉文本特征集合以及视觉特征集合，确定时空候选区域对应的匹配分值，可以包括：

采用如下方式计算匹配分值：

$$s(q,p)=\frac{1}{t_p}\sum_{i=1}^{t_p}s_i\left(h_i^p,h_i^{qp}\right);$$

$$s_i\left(h_i^p,h_i^{qp}\right)=\phi\left(h_i^p,h_i^{qp}\right);$$

其中，$s(q,p)$表示时空候选区域对应的匹配分值，$s_i\left(h_i^p,h_i^{qp}\right)$表示第 i 个时间步数对应的视觉特征和视觉文本特征之间的匹配子分值，h_i^{qp}表示第 i 个时间步数对应的视觉文本特征，h_i^p表示第 i 个时间步数对应的视觉特征，$\phi(\ )$表示相似度计算函数。

本实施例中，介绍了一种计算时空候选区域对应的匹配分值的实现方式。由于一个时空候选区域是由视频序列组成的，一个时空候选区域经过编码处理后会得到一个对应的待匹配视频特征序列，而一个待匹配视频特征序列会对应多个视觉特征，因此，需要针对每个视频特征进行计算，得到匹配子分值，最后每个匹配子分值相加，即可得到整个时空候选区域对应的匹配分值。

为了便于理解，请参阅如下公式：

$$s_i\left(h_i^p,h_i^{qp}\right)=\phi\left(h_i^p,h_i^{qp}\right);$$

这是计算匹配子分值的方式，即计算第 i 个视频特征对应的匹配子分值。类似地，对每个视频特征进行上述计算，即可得到t_p个视频特征的匹配子分值，最后采用如下公式进行计算，即可得到整个时空候选区域的匹配分值：

$$s(q,p)=\frac{1}{t_p}\sum_{i=1}^{t_p}s_i\left(h_i^p,h_i^{qp}\right);$$

更进一步地，本申请实施例还提供了一种根据视觉文本特征集合以及视觉特征集合，确定时空候选区域对应的匹配分值的实现方式。通过上述方式，为方案的实现提供了具体且可行的方式，由此，提升方案的实用性和可行性。

结合上述介绍，下面将对本申请中模型训练的方法进行介绍，请参阅图 6，本申请中模型训练的方法的一个实施例包括：

601、计算机设备获取第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，其中，第一待训练视频与第一待训练文本具有匹配关系，且第一待训练视频与第二待训练文本不具有匹配关系，第二待训练视频与第二待训练文本具有匹配关系，且第二待训练视频与第一待训练文本不具有匹配关系；

本实施例中，模型训练装置首先获取第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，其中，有两对匹配的训练对象以及两对不匹配的训练对象，即第一待训练视频与第一待训练文本匹配，第二待训练视频与第二待训练文本匹配，第一待训练视频与第二待训练文本不匹配，第二待训练视频与第一待训练文本不匹配。

602、计算机设备根据第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，确定排列损失函数，其中，排列损失函数用于对第一待训练视频以及第二待训练文本进行处理，并对第二待训练视频以及第一待训练文本进行处理；

本实施例中，模型训练装置将第一待训练视频以及第一待训练文本作为正样本，将第二待训练视频以及第二待训练文本作为正样本，将第一待训练视频

以及第二待训练文本作为负样本，将第二待训练视频以及第一待训练文本作为负样本，进而获取正样本的匹配分值以及负样本的匹配分值。根据正样本的匹配分值以及负样本的匹配分值之间的大小关系构建排列损失函数。

603、计算机设备根据第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，确定多样性损失函数，其中，多样性损失函数用于对第一待训练视频以及第一待训练文本进行处理，并对第二待训练视频以及第二待训练文本进行处理；

本实施例中，模型训练装置将第一待训练视频以及第一待训练文本作为正样本，将第二待训练视频以及第二待训练文本作为正样本，可以选择任意一个正样本对应的数据构建多样性损失函数。在正样本中，需要使得不同的时空候选区域具有不同的匹配分值，也就是整个时空分布区域的分值分布应该是多样的，而非相等的情况，这样才能实现更为准确的匹配效果。

604、计算机设备根据排列损失函数以及多样性损失函数，确定目标损失函数；

本实施例中，模型训练装置结合排列损失函数以及多样性损失函数，生成目标损失函数。

605、计算机设备采用目标损失函数对待训练的交互器进行训练，得到基于注意力的交互器，其中，该交互器用于输出待匹配视频与待匹配文本的匹配分值。

本实施例中，计算机设备的模型训练装置利用已经构建完成的目标损失函数，对待训练视频文本交互模型进行训练，进而得到视频文本交互模型。将待匹配视频与待匹配文本经过特征化处理之后，输入至视频文本交互模型，由此得到视频中各个时空候选区域的匹配分值，最后选择匹配分值最大的时空候选区域作为目标时空候选区域。

损失函数通常作为学习准则，并且与优化问题相联系，即通过最小化损失函数求解和评估模型。

本申请实施例提供了一种模型训练的方法，首先获取第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，第一待训练视频与第一待训练文本匹配，第一待训练视频与第二待训练文本不匹配，第二待训练视频与第二待训练文本匹配，第二待训练视频与第一待训练文本不匹配。然后根据第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，确定排列损失函数以及多样性损失函数，最后，结合排列损失函数以及多样性损失函数对模型进行训练，得到视频文本交互模型。通过上述方式，同时利用排列损失函数以及多样性损失函数训练模型，能够提升文本与不同时空候选区域之间的匹配准确性，从而有利于提升模型训练的精度。

可选地，在上述图 6 对应的实施例的基础上，本申请实施例提供的模型训练的方法还包括一个可选实施例，在该可选实施例中，上述步骤 602 根据第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，确定排列损失函数，可以包括：

获取第一待训练视频中的第一时空候选区域集合，以及获取第二待训练视频中的第二时空候选区域集合，其中，第一时空候选区域集合包括至少一个第一时空候选区域，第一时空候选区域为视频序列，第二时空候选区域集合包括至少一个第二时空候选区域，第二时空候选区域为视频序列；

根据第一待训练文本以及第二时空候选区域集合，计算第一匹配分值；

根据第二待训练文本以及第一时空候选区域集合，计算第二匹配分值；

根据第一待训练文本以及第一时空候选区域集合，计算第三匹配分值；

根据第一匹配分值、第二匹配分值以及第三匹配分值，确定排列损失函数。

本实施例中，将介绍排列损失函数的内容。首先，模型训练装置对第一待训练视频以及第二待训练视频分别进行时空候选区域的提取，提取方式如上述图 3 对应的第一个可选实施例所描述的内容。基于上述方式，可以得到第一待训练视频对应的第一时空候选区域集合，第一时空候选区域集合包括至少一个第一时空候选区域。并且得到第二待训练视频对应的第二时空候选区域集合，第二时空候选区域集合包括至少一个第二时空候选区域。假设第一待训练视频为 v，第一待训练文本为 q，由此定义二者之间的匹配分值为 $S(v,q)$ 为：

$$S(v,q)=\max_i s(q,p_i), i=1,...,N;$$

其中，p_i 表示第一时空候选区域集合中第 i 个第一时空候选区域，N 表示第一待训练视频中第一时空候选区域的总数，$\max_i(\ )$ 表示取最大值，$s(q,p_i)$ 表示第 i 个第一时空候选区域 p_i 和输入的第一待训练文本 q 之间的匹配行为刻画。$S(v,q)$ 可以表示为第三匹配分值。

类似的，假设第二待训练视频为 v'，第一待训练文本为 q，由此定义二者之间的匹配分值为 $S(v',q)$ 为：

$$S(v',q)=\max_i s(q,p'_i), i=1,...,N;$$

其中，p'_i 表示第二时空候选区域集合中第 i 个第二时空候选区域，N 表示第二待训练视频中第二时空候选区域的总数，$\max_i(\ )$ 表示取最大值，$s(q,p'_i)$ 表示第 i 个第二时空候选区域 p_i 和输入的第一待训练文本 q 之间的匹配行为刻画。$S(v',q)$ 可以表示为第一匹配分值。

类似的，假设第一待训练视频为 v，第二待训练文本为 q'，由此定义二者之间的匹配分值为 $S(v,q')$ 为：

$$S(v,q')=\max_i s(q',p_i), i=1,...,N;$$

其中，p_i 表示第一时空候选区域集合中第 i 个第一时空候选区域，N 表示第一待训练视频中第一时空候选区域的总数，$\max_i(\ )$ 表示取最大值，$s(q',p_i)$ 表示第 i 个第一时空候选区域 p_i 和输入的第二待训练文本 q' 之间的匹配行为刻画。$S(v,q')$ 可以表示为第二匹配分值。

基于上述计算得到的第一匹配分值、第二匹配分值以及第三匹配分值，可以得到如下排列损失函数：

$$L_{rank}=\sum_{v\neq v'}\sum_{q\neq q'}\left[\max\left(0,S(v,q')-S(v,q)+\Delta\right)+\max\left(0,S(v',q)-S(v,q)+\Delta\right)\right];$$

其中，Δ 表示一个常量，排列损失函数 L_{rank} 能够直接促使正样本的第三匹配

分值$S(v,q)$比负样本的第二匹配分值$S(v,q')$以及第一匹配分值$S(v',q)$更大，因此，它能够帮助让目标时空候选区域p^*和第一待训练文本为q之间产生强烈的匹配行为$s(q,p^*)$。

其次，本申请实施例还提供了一种确定排列损失函数的方式，即模型训练装置先获取第一待训练视频中的第一时空候选区域集合，以及获取第二待训练视频中的第二时空候选区域集合，第一时空候选区域集合包括至少一个第一时空候选区域，第一时空候选区域为视频序列，第二时空候选区域集合包括至少一个第二时空候选区域，第二时空候选区域为视频序列，然后分别根据第一待训练文本以及第二时空候选区域集合，计算第一匹配分值，根据第二待训练文本以及第一时空候选区域集合，计算第二匹配分值，根据第一待训练文本以及第一时空候选区域集合，计算第三匹配分值，根据第一匹配分值、第二匹配分值以及第三匹配分值，确定排列损失函数。通过上述方式，设计得到的排列损失函数能够促使匹配数据的匹配分值比未匹配数据的匹配分值更大，从而使得目标时空候选区域和文本之间产生强烈的匹配关系，排列损失函数可以为区分视频和文本之间是否匹配做出贡献。

可选地，在上述图6对应的实施例的基础上，本申请实施例提供的模型训练的方法还包括另一个可选实施例，在该可选实施例中，上述步骤603根据第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，确定多样性损失函数，可以包括：

根据第一时空候选区域集合以及第一待训练文本，确定匹配行为分布，其中，第一时空候选区域集合是根据第一待训练视频生成的，匹配行为分布表示第一时空候选区域集合中每个第一时空候选区域与第一待训练文本之间的匹配关系；对匹配行为分布进行归一化处理，得到目标匹配行为分布；根据目标匹配行为分布确定多样性损失函数。

本实施例中，将介绍多样性损失函数的内容。根据先验经验可知，利用自然语言表达的句子在视频中定位时，仅有很少的一部分时空候选区域是和输入句子有语义配对关系的，这是因为一个合理的匹配行为分布$\{s(q,p_n)\}_{n=1}^{N}$应该是多样性的。也就是，仅有少部分的时空候选区域和文本之间的匹配行为属于强烈匹配行为，另外的时空候选区域的匹配行为应该较弱。

为了使得产生的匹配行为分布$\{s(q,p_n)\}_{n=1}^{N}$具有多样性，由此引入了多样性损失函数。首先，模型训练装置对第一待训练视频进行时空候选区域的提取，提取方式如上述图3对应的第一个可选实施例所描述的内容。基于上述方式，可以得到第一待训练视频对应的时空候选区域集合，时空候选区域集合包括至少一个时空候选区域。假设第一待训练视频为v，第一待训练文本为q。则，本方案先通过softmax函数归一化匹配行为分布$\{s(q,p_n)\}_{n=1}^{N}$，即采用如下方式进行计算：

$$s'(q,p_k)=\frac{\exp(s(q,p_k))}{\sum_{n=1}^{N}\exp(s'(q,p_k))};$$

其中，p_k 是指 p_n 中任意一个，p_k 表示第 k 个时空候选区域，然后再惩罚 $\{s'(q,p_n)\}_{n=1}^{N}$ 分布的信息熵，其作用为增强置信度较高的时空候选区域和文本之间的匹配关系，同时减弱置信度较低的时空候选区域和文本之间的匹配关系。最终得到多样性损失函数：

$$L_{div}=-\sum_{n=1}^{N}s'(q,p_n)\log(s'(q,p_n));$$

其中，L_{div} 表示多样性损失函数。

其次，本申请实施例还提供了一种确定多样性损失函数的方式，即模型训练装置可以根据第一时空候选区域集合以及第一待训练文本，确定匹配行为分布，然后对匹配行为分布进行归一化处理，得到目标匹配行为分布，最后根据目标匹配行为分布确定多样性损失函数。通过上述方式，设计得到的多样性损失函数不仅能够增强置信度较高的时空候选区域和文本之间的匹配关系，同时还能够减弱置信度较低的时空候选区域和文本之间的匹配关系，由此，更贴近时空候选区域与文本之间的实际匹配关系，进而有利于训练得到更加准确的网络模型。

可选地，在上述图 6 以及图 6 对应的两个可选实施例的基础上，本申请实施例提供的模型训练的方法还包括再一个可选实施例，在该可选实施例中，上述步骤 604 根据排列损失函数以及多样性损失函数，确定目标损失函数，可以包括：

采用如下方式确定目标损失函数：

$$L=L_{rank}+\beta L_{div};$$

其中，L 表示目标损失函数，L_{rank} 表示排列损失函数，L_{div} 表示多样性损失函数，β 表示控制系数。

本实施例中，介绍了一种生成目标损失函数的实现方式。在模型训练装置获取到排列损失函数以及多样性损失函数之后，将两者进行相加，同时，为多样性损失函数增加一个系数，详细如下：

$$L=L_{rank}+\beta L_{div};$$

其中，β 可以设置为 0.5，也可以设置为其他合理的数值，此处不做限定。

即，计算机设备获取控制系数，根据所述控制系数、所述排列损失函数以及所述多样性损失函数，确定所述目标损失函数。

再次，本申请实施例还提供了一种确定目标损失函数的实现方式，即对已经设计好的排列损失函数与多样性损失函数进行组合。通过上述方式，设计得到的目标损失函数既可以区分匹配的视频和句子对以及不匹配的视频和句子对，又可以增强匹配的视频和句子对中可信度高的时空候选区域和句子的匹配关系，并且能够降低可信度低的时空候选区域和句子之间的匹配关系，由此，提升模型训练的可靠性，从而得到更为准确的网络模型。

下面对本申请中的视频序列选择装置进行详细描述，请参阅图 7，图 7 为本申请实施例中视频序列选择装置一个实施例示意图，视频序列选择装置 70 包括：

获取模块 701，用于接收待匹配视频以及待匹配文本，其中，所述待匹配视频包括多帧图像，所述待匹配文本包括至少一个词语，所述待匹配文本对应于待匹配文本特征序列；

生成模块 702，用于调用时空候选区域生成器从所述待匹配视频中提取时空候选区域集合，其中，所述时空候选区域集合中包括 N 个时空候选区域，所述 N 为大于或等于 1 的整数，一个时空候选区域为一个视频序列；

编码模块 703，用于通过卷积神经网络对所述时空候选区域集合中的每个时空候选区域进行特征提取，得到 N 个待匹配视频特征序列，其中，所述待匹配视频特征序列与所述时空候选区域具有对应关系；

所述获取模块 701，还用于调用基于注意力的交互器所述每个时空候选区域对应的匹配分值，其中，所述交互器用于对所述待匹配视频特征序列与所述待匹配文本特征序列进行处理，所述匹配分值用于表示所述时空候选区域与所述待匹配文本之间的匹配关系；

选择模块 704，用于根据交互器输出所述每个时空候选区域对应的匹配分值，从所述时空候选区域集合中选择目标时空候选区域，输出所述目标时空候选区域。

本申请实施例提供了一种视频序列选择装置，首先接收待匹配视频以及待匹配文本，其中，待匹配视频包括多帧图像，待匹配文本包括至少一个词语，待匹配文本对应于待匹配文本特征序列，然后从待匹配视频中提取时空候选区域集合，接下来需要对时空候选区域集合中的每个时空候选区域进行特征提取，得到 N 个待匹配视频特征序列，其中，待匹配视频特征序列与时空候选区域具有对应关系，然后可以调用基于注意力的交互器获取每个时空候选区域对应的匹配分值，最后根据每个时空候选区域对应的匹配分值，从时空候选区域集合中选择目标时空候选区域，其中，一个时空候选区域为一个视频序列。通过上述方式，对视频中的时空候选区域和文本进行匹配，而不再是对视频中的每帧图像与文本进行匹配，这样操作的好处是，由于时空候选区域包括了图像在时间和空间上的关系，因此，在匹配的时候考虑到了视频与文本在时序上的关联性，即考虑了视频时序信息对视频序列以及文本的影响，从而提升了输出的视频序列与文本的匹配度，进而有利于更好地理解视频内容。

可选地，在上述图 7 对应的实施例的基础上，本申请实施例提供的视频序列选择装置 70 的另一个可选实施例中，

所述生成模块 702，用于调用所述时空候选区域生成器获取所述待匹配视频中每帧图像的候选区域以及置信度得分，其中，每个候选区域对应一个置信度得分；获取所述待匹配视频中相邻两帧图像之间的重合度；根据所述每帧图像的候选区域、所述置信度得分以及所述重合度，生成所述时空候选区域集合。

可选地，在上述图 7 对应的实施例的基础上，本申请实施例提供的视频序列选择装置 70 的另一个可选实施例中，

所述获取模块 701，用于对于所述每个时空候选区域，调用所述交互器的编码器对所述时空候选区域对应的待匹配视频特征序列进行编码，得到视觉特征集合，其中，所述视觉特征集合包括至少一个视觉特征；调用所述交互器的编码器对所述待匹配文本特征序列进行编码，得到文本特征集合，其中，所述文本特征集合包括至少一个文本特征；根据所述视觉特征集合以及所述文本特征集合，确定视觉文本特征集合，其中，所述视觉文本特征集合包括至少一个视觉文本特征，所述视觉文本特征表示基于视觉特征的文本特征；根据所述视觉文本特征集合以及所述视觉特征集合，确定所述时空候选区域对应的匹配分值。

可选地，在上述图 7 对应的实施例的基础上，本申请实施例提供的视频序列选择装置 70 的另一个可选实施例中，

所述获取模块 701，用于采用如下方式计算所述视觉特征集合：

$$H_p = \left\{ h_t^p \right\}_{t=1}^{t_p};$$

$$h_t^p = LSTM_p\left(f_t^p, h_{t-1}^p \right);$$

其中，所述 H_p 表示所述视觉特征集合，所述 h_t^p 表示所述视觉特征集合中的第 t 个视觉特征，所述 t_p 表示所述时空候选区域的时间步数，所述 h_{t-1}^p 表示所述视觉特征集合中的第（t-1）个视觉特征，所述 $LSTM_p(\ )$ 表示第一长短期记忆网络 LSTM 编码器，所述 f_t^p 表示所述待匹配视频特征序列中的第 t 行特征；

采用如下方式计算所述文本特征集合：

$$H_q = \left\{ h_t^q \right\}_{t=1}^{t_q};$$

$$h_t^q = LSTM_q\left(f_t^q, h_{t-1}^q \right);$$

其中，所述 H_q 表示所述文本特征集合，所述 h_t^q 表示所述文本特征集合中的第 t 个文本特征，所述 t_q 表示所述待匹配文本的词语数量，所述 h_{t-1}^q 表示所述文本特征集合中的第（t-1）个文本特征，所述 $LSTM_q(\ )$ 表示第二 LSTM 编码器，所述 f_t^q 表示所述待匹配文本特征序列中的第 t 行特征。

可选地，在上述图 7 对应的实施例的基础上，本申请实施例提供的视频序列选择装置 70 的另一个可选实施例中，

所述获取模块 701，用于调用所述交互器执行根据所述视觉特征集合以及所述文本特征集合，获取得到视觉特征对应文本特征的注意力权重；根据所述注意力权重，获取所述视觉特征对应所述文本特征的归一化注意力权重；根据所述归一化注意力权重以及所述文本特征，获取视觉文本特征集合。

可选地，在上述图 7 对应的实施例的基础上，本申请实施例提供的视频序列选择装置 70 的另一个可选实施例中，

所述获取模块 701，用于采用如下方式计算所述注意力权重：

$e_{i,j} = w^T \tanh\left(W^q h_j^q + W^p h_i^p + b_1\right) + b_2$；

其中，所述$e_{i,j}$表示第 i 个视觉特征对应第 j 个文本特征的注意力权重，所述h_j^q表示所述第 j 个文本特征，所述h_i^p表示所述第 i 个视觉特征，所述w^T表示第一模型参数，所述W^q表示第二模型参数，所述W^p表示第三模型参数，所述b_1表示第四模型参数，所述b_2表示第五模型参数，所述$\tanh(\ )$表示双曲正切函数；

采用如下方式计算所述归一化注意力权重：

$$a_{i,j} = \frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{t_q} \exp\left(e_{i,k}\right)};$$

其中，所述$a_{i,j}$表示所述第 i 个视觉特征对应所述第 j 个文本特征的归一化注意力权重，所述t_q表示所述待匹配文本的词语数量，所述k表示所述待匹配文本中的第k个词语，所述k大于或等于 1，且小于或等于所述t_q的整数，所述$\exp(\ )$表示指数函数；

采用如下方式计算所述视觉文本特征集合：

$$H_{qp} = \left\{h_{qp}\right\}_{t=1}^{t_p};$$

$$h_{qp} = \sum_{j=1}^{t_q} a_{i,j} h_j^q;$$

其中，所述H_{qp}表示所述视觉文本特征集合，所述t_p表示所述时空候选区域的时间步数，所述h_{qp}表示视觉文本特征。

可选地，在上述图 7 对应的实施例的基础上，本申请实施例提供的视频序列选择装置 70 的另一个可选实施例中，

所述获取模块 701，用于采用如下方式计算所述匹配分值：

$$s(q,p) = \frac{1}{t_p} \sum_{i=1}^{t_p} s_i\left(h_i^p, h_i^{qp}\right);$$

$$s_i\left(h_i^p, h_i^{qp}\right) = \phi\left(h_i^p, h_i^{qp}\right);$$

其中，所述$s(q,p)$表示所述时空候选区域对应的匹配分值，所述$s_i\left(h_i^p, h_i^{qp}\right)$表示第 i 个时间步数对应的视觉特征和视觉文本特征之间的匹配子分值，所述h_i^{qp}表示所述第 i 个时间步数对应的视觉文本特征，所述h_i^p表示所述第 i 个时间步数对应的视觉特征，所述$\phi(\ )$表示相似度计算函数。

下面对本申请中的模型训练装置进行详细描述，请参阅图 8，图 8 为本申请实施例中模型训练装置一个实施例示意图，模型训练装置 80 包括：

获取模块 801，用于获取第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，其中，所述第一待训练视频与所述第一待训练文本具

有匹配关系，且所述第一待训练视频与所述第二待训练文本不具有匹配关系，所述第二待训练视频与所述第二待训练文本具有匹配关系，且所述第二待训练视频与所述第一待训练文本不具有匹配关系；

确定模块802，用于根据所述获取模块801获取的所述第一待训练视频、所述第二待训练视频、所述第一待训练文本以及所述第二待训练文本，确定排列损失函数，其中，所述排列损失函数用于对所述第一待训练视频以及所述第二待训练文本进行处理，并对所述第二待训练视频以及所述第一待训练文本进行处理；

所述确定模块802，还用于根据所述获取模块801获取的所述第一待训练视频、所述第二待训练视频、所述第一待训练文本以及所述第二待训练文本，确定多样性损失函数，其中，所述多样性损失函数用于对所述第一待训练视频以及所述第一待训练文本进行处理，并对所述第二待训练视频以及所述第二待训练文本进行处理；

所述确定模块802，还用于根据所述排列损失函数以及所述多样性损失函数，确定目标损失函数；

训练模块803，用于采用所述确定模块802确定的所述目标损失函数对待训练的交互器进行训练，得到基于注意力的交互器，其中，所述交互器用于输出待匹配视频与待匹配文本的匹配分值。

本申请实施例中，提供了一种模型训练装置，首先获取第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，第一待训练视频与第一待训练文本匹配，第一待训练视频与第二待训练文本不匹配，第二待训练视频与第二待训练文本匹配，第二待训练视频与第一待训练文本不匹配。然后根据第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，确定排列损失函数以及多样性损失函数，最后，结合排列损失函数以及多样性损失函数对模型进行训练，得到基于注意力的交互器。通过上述方式，同时利用排列损失函数以及多样性损失函数训练模型，不仅能够提升文本与不同时空候选区域之间的匹配准确性，还可以提升文本与候选区域之间的匹配准确性，从而有利于提升模型训练的精度。

可选地，在上述图8对应的实施例的基础上，本申请实施例提供的模型训练装置80的另一个可选实施例中，

所述确定模块802，用于获取所述第一待训练视频中的第一时空候选区域集合，以及获取所述第二待训练视频中的第二时空候选区域集合，其中，所述第一时空候选区域集合包括至少一个第一时空候选区域，所述第一时空候选区域为视频序列，所述第二时空候选区域集合包括至少一个第二时空候选区域，所述第二时空候选区域为视频序列；根据所述第一待训练文本以及所述第二时空候选区域集合，计算第一匹配分值；根据所述第二待训练文本以及所述第一时空候选区域集合，计算第二匹配分值；根据所述第一待训练文本以及所述第一时空候选区域集合，计算第三匹配分值；根据所述第一匹配分值、所述第二匹配分值以及所述第三匹配分值，确定所述排列损失函数。

可选地，在上述图8对应的实施例的基础上，本申请实施例提供的模型训

练装置 80 的另一个可选实施例中，

所述确定模块 802，用于根据第一时空候选区域集合以及所述第一待训练文本，确定匹配行为分布，其中，所述第一时空候选区域集合是根据所述第一待训练视频生成的，所述匹配行为分布表示所述第一时空候选区域集合中每个第一时空候选区域与所述第一待训练文本之间的匹配关系；对所述匹配行为分布进行归一化处理，得到目标匹配行为分布；根据所述目标匹配行为分布确定所述多样性损失函数。

可选地，在上述图 8 对应的实施例的基础上，本申请实施例提供的模型训练装置 80 的另一个可选实施例中，所述确定模块 802，用于获取控制系数，根据所述控制系数、所述排列损失函数以及所述多样性损失函数，确定所述目标损失函数。

图 9 是本发明实施例提供的一种服务器结构示意图，该服务器 900 可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上中央处理器（central processing units，CPU）922（例如，一个或一个以上处理器）和存储器 932，一个或一个以上存储应用程序 942 或数据 944 的存储介质 930（例如一个或一个以上海量存储设备）。其中，存储器 932 和存储介质 930 可以是短暂存储或持久存储。存储在存储介质 930 的程序可以包括一个或一个以上模块（图示没标出），每个模块可以包括对服务器中的一系列指令操作。更进一步地，中央处理器 922 可以设置为与存储介质 930 通信，在服务器 900 上执行存储介质 530 中的一系列指令操作。

服务器 900 还可以包括一个或一个以上电源 926，一个或一个以上有线或无线网络接口 950，一个或一个以上输入输出接口 958，和/或，一个或一个以上操作系统 941，例如 Windows ServerTM，Mac OS XTM，UnixTM，LinuxTM，FreeBSDTM 等等。

上述实施例中由服务器所执行的步骤可以基于该图 9 所示的服务器结构。

本申请实施例中，服务器中的 CPU 922 用于执行上述实施例中提供的视频序列选择的方法或模型训练的方法。

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。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

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单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，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）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器（read-only memory，ROM）、随机存取存储器（random access memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，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申请各实施例技术方案的精神和范围。

权 利 要 求 书

1、一种视频序列选择的方法，其特征在于，所述方法应用于计算机设备，包括：

所述计算机设备接收待匹配视频以及待匹配文本，其中，所述待匹配视频包括多帧图像，所述待匹配文本包括至少一个词语，所述待匹配文本对应于待匹配文本特征序列；

所述计算机设备调用时空候选区域生成器从所述待匹配视频中提取时空候选区域集合，其中，所述时空候选区域集合中包括 N 个时空候选区域，所述 N 为大于或等于 1 的整数，一个时空候选区域为一个视频序列；

所述计算机设备通过卷积神经网络对所述时空候选区域集合中的每个时空候选区域进行特征提取，得到 N 个待匹配视频特征序列，其中，所述待匹配视频特征序列与所述时空候选区域具有对应关系；

所述计算机设备调用基于注意力的交互器获取所述每个时空候选区域对应的匹配分值，其中，所述交互器用于对所述待匹配视频特征序列与所述待匹配文本特征序列进行处理，所述匹配分值用于表示所述时空候选区域与所述待匹配文本之间的匹配关系；

所述计算机设备根据所述交互器输出的所述每个时空候选区域对应的匹配分值，从所述时空候选区域集合中选择目标时空候选区域，输出所述目标时空候选区域。

2、根据权利要求 1 所述的方法，其特征在于，所述计算机设备调用时空候选区域生成器从所述待匹配视频中提取时空候选区域集合，包括：

所述计算机设备调用所述时空候选区域生成器获取所述待匹配视频中每帧图像的候选区域以及置信度得分，其中，每个候选区域对应一个置信度得分；

所述计算机设备调用所述时空候选区域生成器获取所述待匹配视频中相邻两帧图像之间的重合度；

所述计算机设备调用所述时空候选区域生成器执行根据所述每帧图像的候选区域、所述置信度得分以及所述重合度，生成所述时空候选区域集合。

3、根据权利要求 1 所述的方法，其特征在于，所述计算机设备调用基于注意力的交互器获取所述每个时空候选区域对应的匹配分值，包括：

对于所述每个时空候选区域，所述计算机设备调用所述交互器的编码器对所述时空候选区域对应的待匹配视频特征序列进行编码，得到视觉特征集合，其中，所述视觉特征集合包括至少一个视觉特征；

所述计算机设备调用所述交互器的编码器对所述待匹配文本特征序列进行编码，得到文本特征集合，其中，所述文本特征集合包括至少一个文本特征；

所述计算机设备调用所述交互器执行根据所述视觉特征集合以及所述文本

特征集合，确定视觉文本特征集合，其中，所述视觉文本特征集合包括至少一个视觉文本特征，所述视觉文本特征表示基于视觉特征的文本特征；

所述计算机设备调用所述交互器执行根据所述视觉文本特征集合以及所述视觉特征集合，确定所述时空候选区域对应的匹配分值。

4、根据权利要求 3 所述的方法，其特征在于，所述计算机设备调用所述交互器的编码器对所述时空候选区域对应的待匹配视频特征序列进行编码，得到视觉特征集合，包括：

采用如下方式计算所述视觉特征集合：

$$H_p = \left\{ h_t^p \right\}_{t=1}^{t_p};$$

$$h_t^p = LSTM_p\left(f_t^p, h_{t-1}^p \right);$$

其中，所述 H_p 表示所述视觉特征集合，所述 h_t^p 表示所述视觉特征集合中的第 t 个视觉特征，所述 t_p 表示所述时空候选区域的时间步数，所述 h_{t-1}^p 表示所述视觉特征集合中的第（t-1）个视觉特征，所述 $LSTM_p(\ )$ 表示第一长短期记忆网络 LSTM 编码器，所述 f_t^p 表示所述待匹配视频特征序列中的第 t 行特征；

所述计算机设备调用所述交互器的编码器对所述待匹配文本特征序列进行编码，得到文本特征集合，包括：

采用如下方式计算所述文本特征集合：

$$H_q = \left\{ h_t^q \right\}_{t=1}^{t_q};$$

$$h_t^q = LSTM_q\left(f_t^q, h_{t-1}^q \right);$$

其中，所述 H_q 表示所述文本特征集合，所述 h_t^q 表示所述文本特征集合中的第 t 个文本特征，所述 t_q 表示所述待匹配文本的词语数量，所述 h_{t-1}^q 表示所述文本特征集合中的第（t-1）个文本特征，所述 $LSTM_q(\ )$ 表示第二 LSTM 编码器，所述 f_t^q 表示所述待匹配文本特征序列中的第 t 行特征。

5、根据权利要求 3 所述的方法，其特征在于，所述计算机设备调用所述交互器执行根据所述视觉特征集合以及所述文本特征集合，确定视觉文本特征集合，包括：

所述计算机设备调用所述交互器执行根据所述视觉特征集合以及所述文本特征集合，计算视觉特征对应文本特征的注意力权重；

所述计算机设备调用所述交互器执行根据所述注意力权重，计算所述视觉特征对应所述文本特征的归一化注意力权重；

所述计算机设备调用所述交互器执行根据所述归一化注意力权重以及所述文本特征，计算视觉文本特征集合。

6、根据权利要求 5 所述的方法，其特征在于，所述计算机设备调用所述交互器执行根据所述视觉特征集合以及所述文本特征集合，计算视觉特征对应文本特征的注意力权重，包括：

采用如下方式计算所述注意力权重：

$$e_{i,j} = w^T \tanh\left(W^q h_j^q + W^p h_i^p + b_1\right) + b_2 ;$$

其中，所述$e_{i,j}$表示第 i 个视觉特征对应第 j 个文本特征的注意力权重，所述h_j^q表示所述第 j 个文本特征，所述h_i^p表示所述第 i 个视觉特征，所述w^T表示第一模型参数，所述W^q表示第二模型参数，所述W^p表示第三模型参数，所述b_1表示第四模型参数，所述b_2表示第五模型参数，所述$\tanh(\ )$表示双曲正切函数；

所述计算机设备调用所述交互器执行根据所述注意力权重，计算所述视觉特征对应所述文本特征的归一化注意力权重，包括：

采用如下方式计算所述归一化注意力权重：

$$a_{i,j} = \frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{t_q} \exp\left(e_{i,k}\right)} ;$$

其中，所述$a_{i,j}$表示所述第 i 个视觉特征对应所述第 j 个文本特征的归一化注意力权重，所述t_q表示所述待匹配文本的词语数量，所述k表示所述待匹配文本中的第k个词语，所述k为大于或等于 1，且小于或等于所述t_q的整数，所述$\exp(\ )$表示指数函数；

所述计算机设备调用所述交互器执行根据所述归一化注意力权重以及所述文本特征，计算视觉文本特征集合，包括：

采用如下方式计算所述视觉文本特征集合：

$$H_{qp} = \left\{h_{qp}\right\}_{t=1}^{t_p} ;$$

$$h_{qp} = \sum_{j=1}^{t_q} a_{i,j} h_j^q ;$$

其中，所述H_{qp}表示所述视觉文本特征集合，所述t_p表示所述时空候选区域的时间步数，所述h_{qp}表示视觉文本特征。

7、根据权利要求 3 至 6 中任一项所述的方法，其特征在于，所述计算机设备调用所述交互器执行根据所述视觉文本特征集合以及所述视觉特征集合，确定所述时空候选区域对应的匹配分值，包括：

采用如下方式计算所述匹配分值：

$$s(q,p) = \frac{1}{t_p} \sum_{i=1}^{t_p} s_i\left(h_i^p, h_i^{qp}\right) ;$$

$$s_i\left(h_i^p, h_i^{qp}\right) = \phi\left(h_i^p, h_i^{qp}\right) ;$$

其中，所述$s(q,p)$表示所述时空候选区域对应的匹配分值，所述$s_i\left(h_i^p,h_i^{qp}\right)$表示第 i 个时间步数对应的视觉特征和视觉文本特征之间的匹配子分值，所述h_i^{qp}表示所述第 i 个时间步数对应的视觉文本特征，所述h_i^p表示所述第 i 个时间步数对应的视觉特征，所述$\phi(\ )$表示相似度获取函数。

8、根据权利要求 1 所述的方法，其特征在于，还包括：

所述计算机设备获取第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，其中，所述第一待训练视频与所述第一待训练文本具有匹配关系，且所述第一待训练视频与所述第二待训练文本不具有匹配关系，所述第二待训练视频与所述第二待训练文本具有匹配关系，且所述第二待训练视频与所述第一待训练文本不具有匹配关系；

所述计算机设备根据所述第一待训练视频、所述第二待训练视频、所述第一待训练文本以及所述第二待训练文本，确定排列损失函数，其中，所述排列损失函数用于对所述第一待训练视频以及所述第二待训练文本进行处理，并对所述第二待训练视频以及所述第一待训练文本进行处理；

所述计算机设备根据所述第一待训练视频、所述第二待训练视频、所述第一待训练文本以及所述第二待训练文本，确定多样性损失函数，其中，所述多样性损失函数用于对所述第一待训练视频以及所述第一待训练文本进行处理，并对所述第二待训练视频以及所述第二待训练文本进行处理；

所述计算机设备根据所述排列损失函数以及所述多样性损失函数，确定目标损失函数；

所述计算机设备采用所述目标损失函数对待训练的交互器进行训练，得到所述交互器，其中，所述交互器用于输出待匹配视频与待匹配文本的匹配分值。

9、根据权利要求 8 所述的方法，其特征在于，所述计算机设备根据所述第一待训练视频、所述第二待训练视频、所述第一待训练文本以及所述第二待训练文本，确定排列损失函数，包括：

所述计算机设备获取所述第一待训练视频中的第一时空候选区域集合，以及获取所述第二待训练视频中的第二时空候选区域集合，其中，所述第一时空候选区域集合包括至少一个第一时空候选区域，所述第一时空候选区域为视频序列，所述第二时空候选区域集合包括至少一个第二时空候选区域，所述第二时空候选区域为视频序列；

所述计算机设备根据所述第一待训练文本以及所述第二时空候选区域集合，计算第一匹配分值；

所述计算机设备根据所述第二待训练文本以及所述第一时空候选区域集合，计算第二匹配分值；

所述计算机设备根据所述第一待训练文本以及所述第一时空候选区域集合，计算第三匹配分值；

所述计算机设备根据所述第一匹配分值、所述第二匹配分值以及所述第三匹配分值，确定所述排列损失函数。

10、根据权利要求 8 所述的方法，其特征在于，所述计算机设备根据所述第一待训练视频、所述第二待训练视频、所述第一待训练文本以及所述第二待训练文本，确定多样性损失函数，包括：

所述计算机设备根据第一时空候选区域集合以及所述第一待训练文本，确定匹配行为分布，其中，所述第一时空候选区域集合是根据所述第一待训练视频生成的，所述匹配行为分布表示所述第一时空候选区域集合中每个第一时空候选区域与所述第一待训练文本之间的匹配关系；

所述计算机设备对所述匹配行为分布进行归一化处理，得到目标匹配行为分布；

所述计算机设备根据所述目标匹配行为分布确定所述多样性损失函数。

11、根据权利要求 8 至 10 中任一项所述的方法，其特征在于，所述计算机设备根据所述排列损失函数以及所述多样性损失函数，确定目标损失函数，包括：

所述计算机设备获取控制系数，根据所述控制系数、所述排列损失函数以及所述多样性损失函数，确定所述目标损失函数。

12、一种计算机设备，其特征在于，包括：存储器、收发器、处理器以及总线系统；

其中，所述存储器用于存储程序；

所述处理器用于执行所述存储器中的程序，包括如下步骤：

接收待匹配视频以及待匹配文本，其中，所述待匹配视频包括多帧图像，所述待匹配文本包括至少一个词语，所述待匹配文本对应于待匹配文本特征序列；

调用时空候选区域生成器从所述待匹配视频中提取时空候选区域集合，其中，所述时空候选区域集合中包括 N 个时空候选区域，所述 N 为大于或等于 1 的整数，一个时空候选区域为一个视频序列；

通过卷积神经网络对所述时空候选区域集合中的每个时空候选区域进行特征提取，得到 N 个待匹配视频特征序列，其中，所述待匹配视频特征序列与所述时空候选区域具有对应关系；

调用基于注意力的交互器获取所述每个时空候选区域对应的匹配分值，其中，所述交互器用于对所述待匹配视频特征序列与所述待匹配文本特征序列进行处理，所述匹配分值用于表示所述时空候选区域与所述待匹配文本之间的匹配关系；

根据所述交互器输出的所述每个时空候选区域对应的匹配分值，从所述时空候选区域集合中选择目标时空候选区域，输出所述目标时空候选区域；

所述总线系统用于连接所述存储器以及所述处理器，以使所述存储器以及所述处理器进行通信。

13、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有指令，当其在计算机设备上运行时，使得计算机设备执行如权利要求 1 至 11 中任一项所述的视频序列选择的方法。

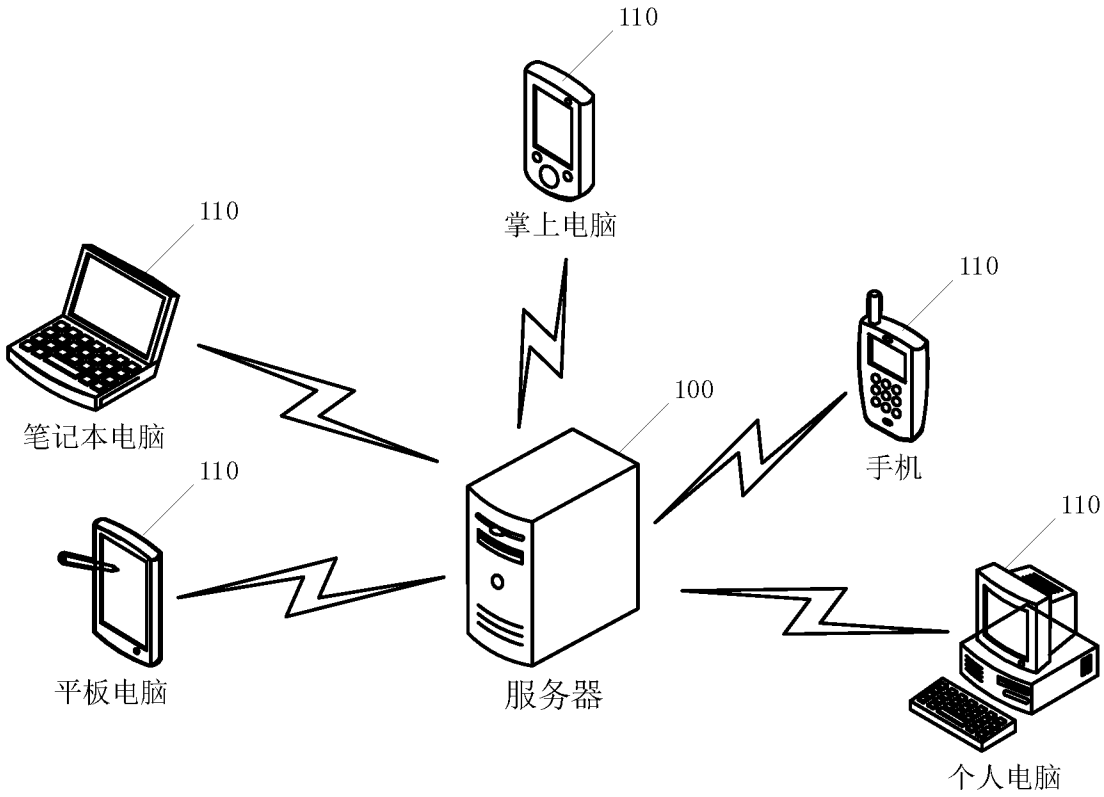
110
掌上电脑
110
笔记本电脑
110
手机
100
110
平板电脑
服务器
110
个人电脑

图 1

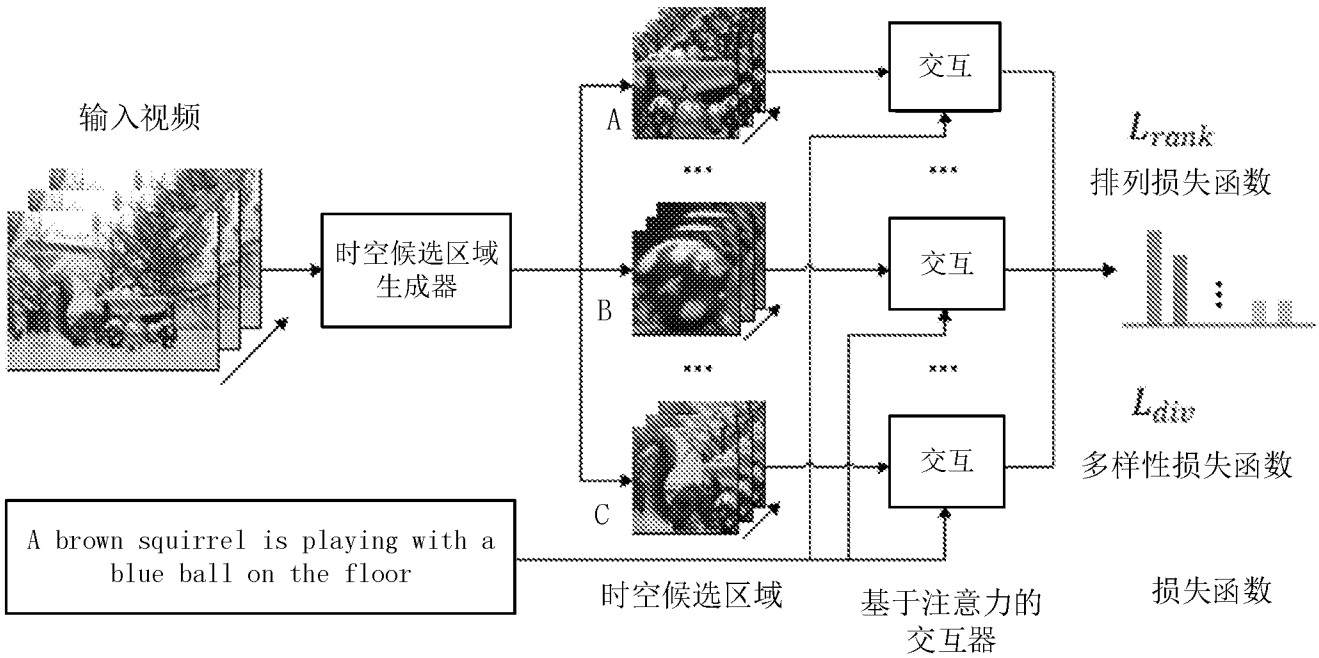
输入视频
时空候选区域
生成器
A
B
C
交互
交互
交互
L_{rank}
排列损失函数
L_{div}
多样性损失函数
A brown squirrel is playing with a blue ball on the floor
时空候选区域
基于注意力的
交互器
损失函数

图 2

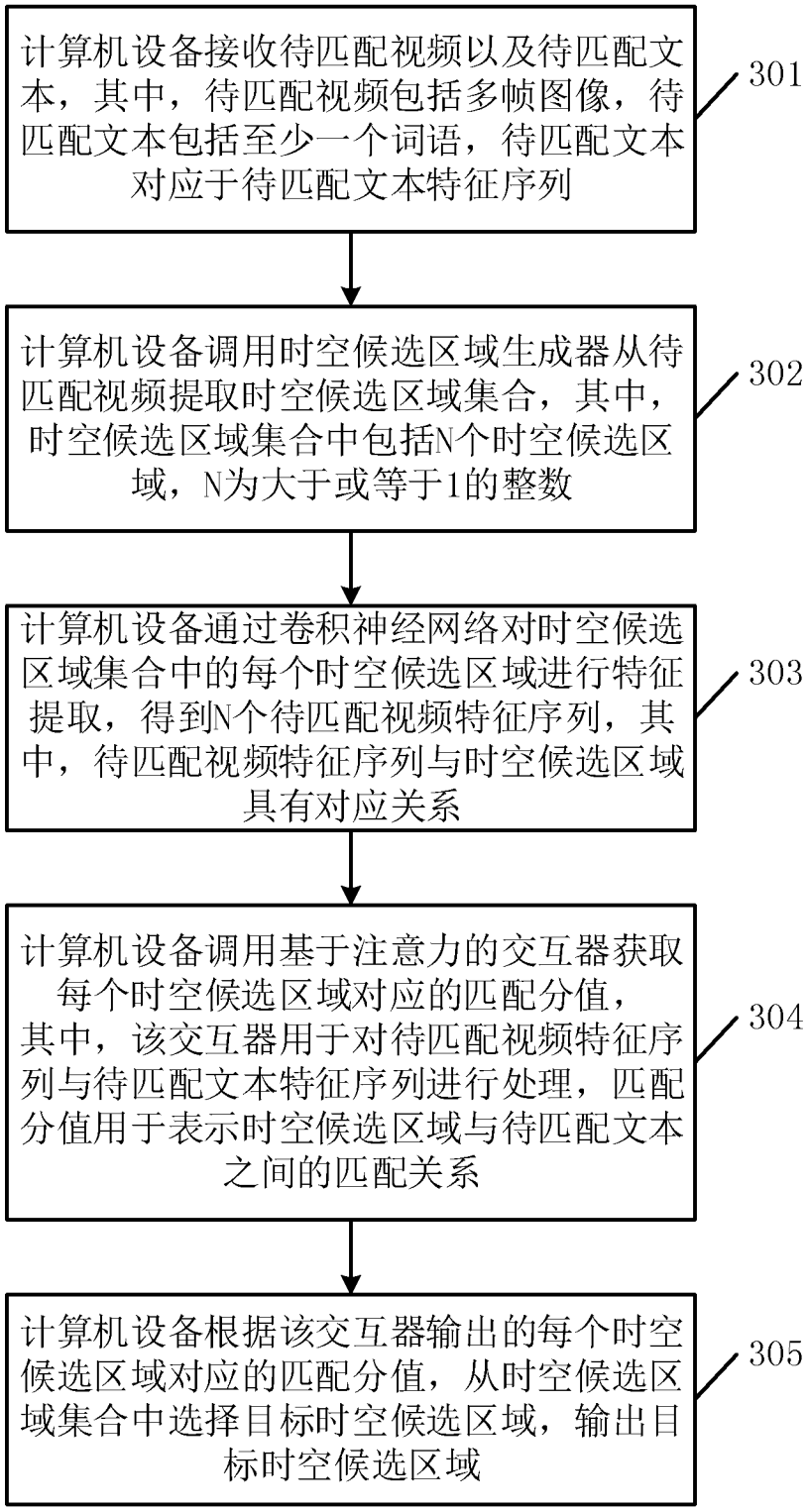
计算机设备接收待匹配视频以及待匹配文本，其中，待匹配视频包括多帧图像，待匹配文本包括至少一个词语，待匹配文本对应于待匹配文本特征序列
301
计算机设备调用时空候选区域生成器从待匹配视频提取时空候选区域集合，其中，时空候选区域集合中包括N个时空候选区域，N为大于或等于1的整数
302
计算机设备通过卷积神经网络对时空候选区域集合中的每个时空候选区域进行特征提取，得到N个待匹配视频特征序列，其中，待匹配视频特征序列与时空候选区域具有对应关系
303
计算机设备调用基于注意力的交互器获取每个时空候选区域对应的匹配分值，其中，该交互器用于对待匹配视频特征序列与待匹配文本特征序列进行处理，匹配分值用于表示时空候选区域与待匹配文本之间的匹配关系
304
计算机设备根据该交互器输出的每个时空候选区域对应的匹配分值，从时空候选区域集合中选择目标时空候选区域，输出目标时空候选区域
305

图 3

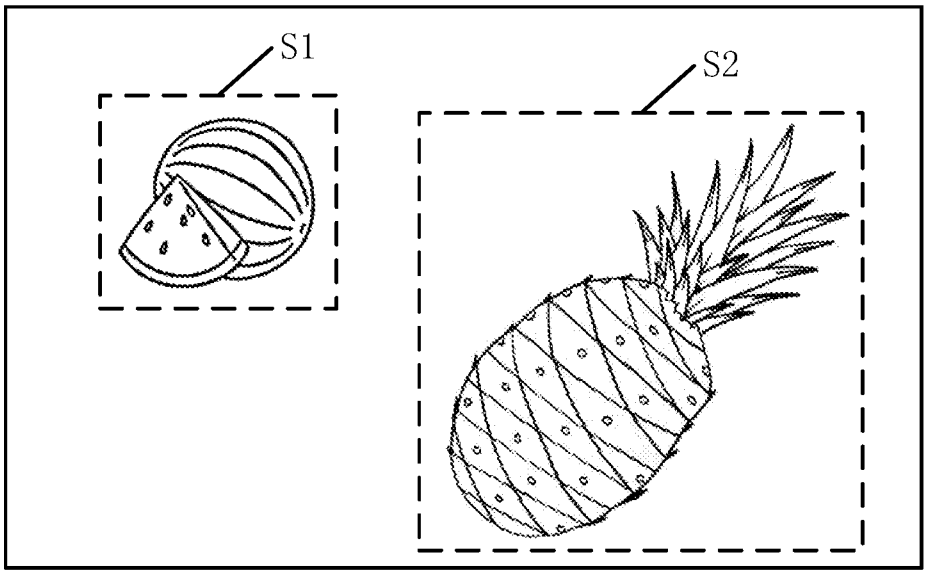
S1
S2

图 4

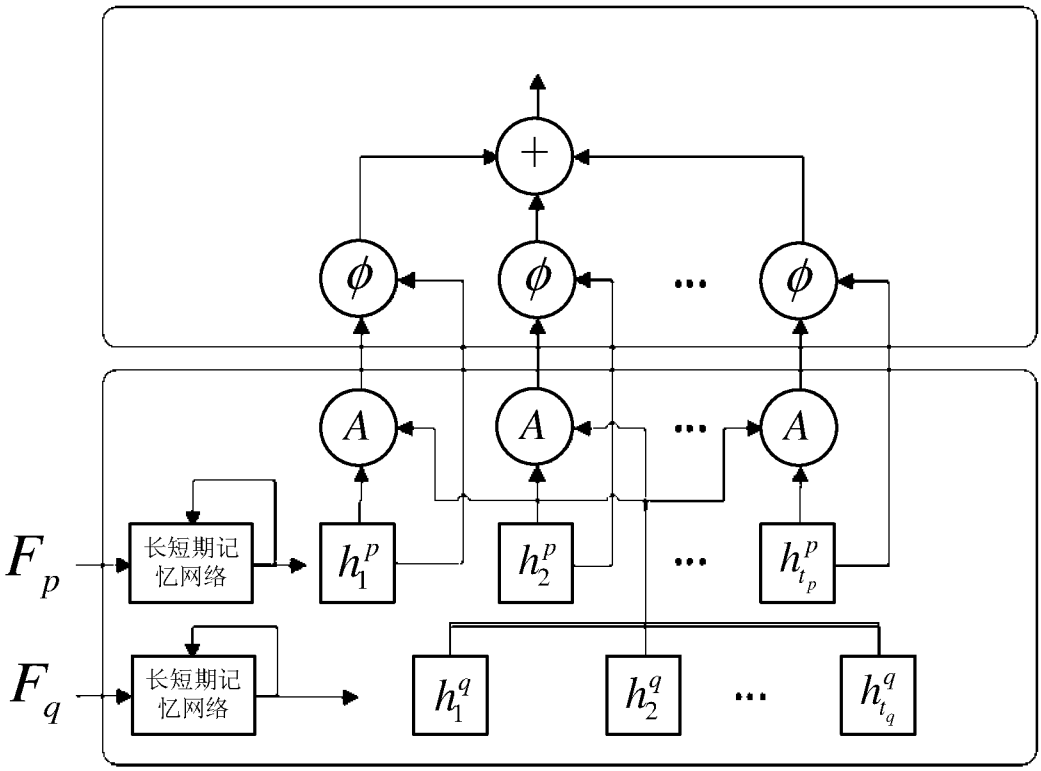
F_p
F_q
长短期记忆网络
长短期记忆网络
h_1^p
h_2^p
$h_{t_p}^p$
h_1^q
h_2^q
$h_{t_q}^q$
A
A
A
ϕ
ϕ
ϕ
+

图 5

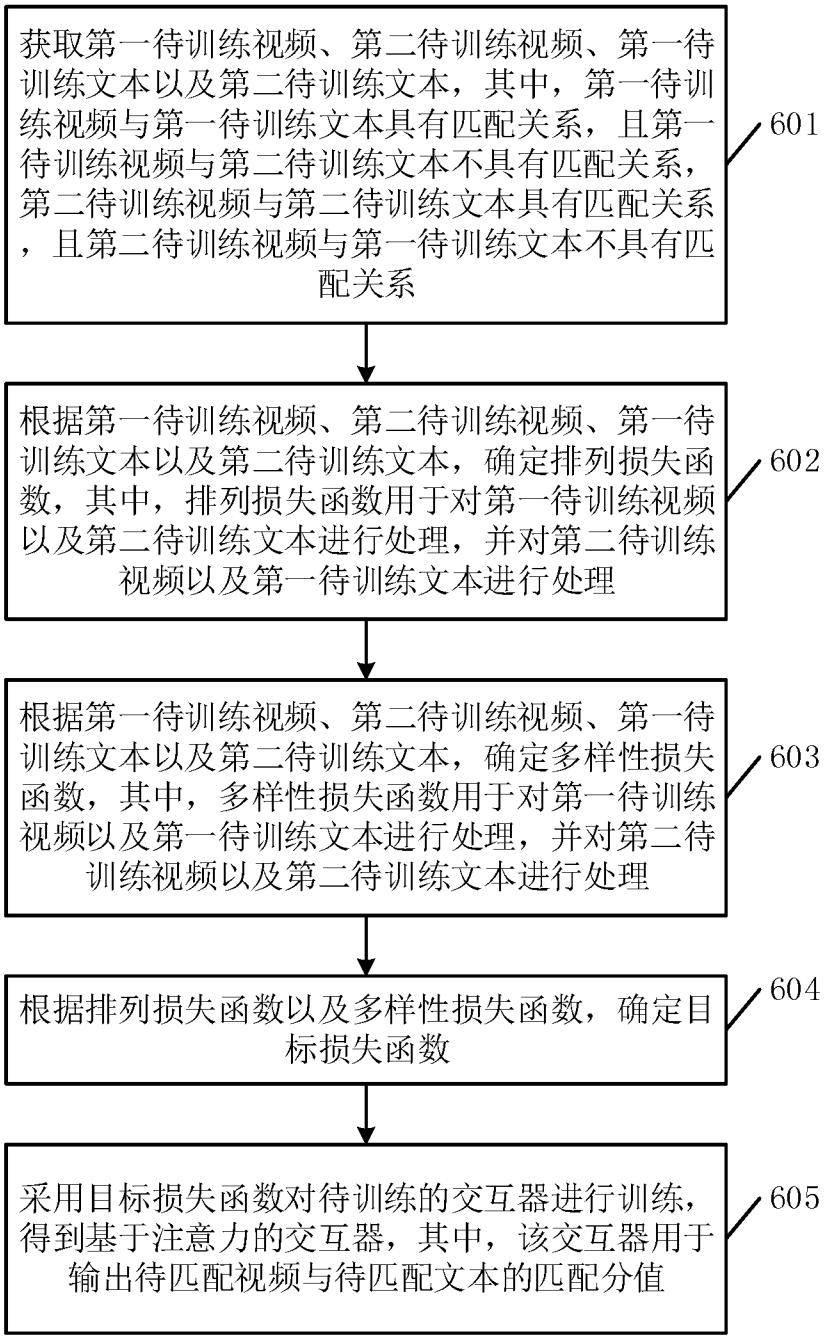
获取第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，其中，第一待训练视频与第一待训练文本具有匹配关系，且第一待训练视频与第二待训练文本不具有匹配关系，第二待训练视频与第二待训练文本具有匹配关系，且第二待训练视频与第一待训练文本不具有匹配关系
601
根据第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，确定排列损失函数，其中，排列损失函数用于对第一待训练视频以及第二待训练文本进行处理，并对第二待训练视频以及第一待训练文本进行处理
602
根据第一待训练视频、第二待训练视频、第一待训练文本以及第二待训练文本，确定多样性损失函数，其中，多样性损失函数用于对第一待训练视频以及第一待训练文本进行处理，并对第二待训练视频以及第二待训练文本进行处理
603
根据排列损失函数以及多样性损失函数，确定目标损失函数
604
采用目标损失函数对待训练的交互器进行训练，得到基于注意力的交互器，其中，该交互器用于输出待匹配视频与待匹配文本的匹配分值
605

图 6

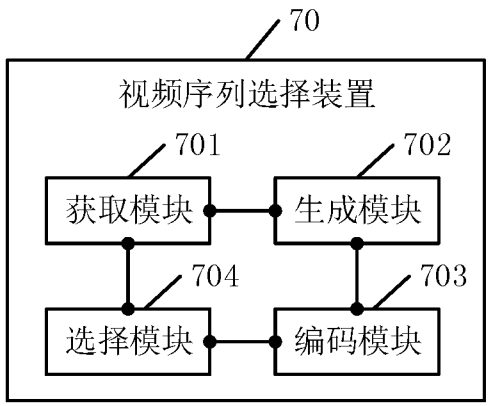
70
视频序列选择装置
701
获取模块
702
生成模块
704
选择模块
703
编码模块

图 7

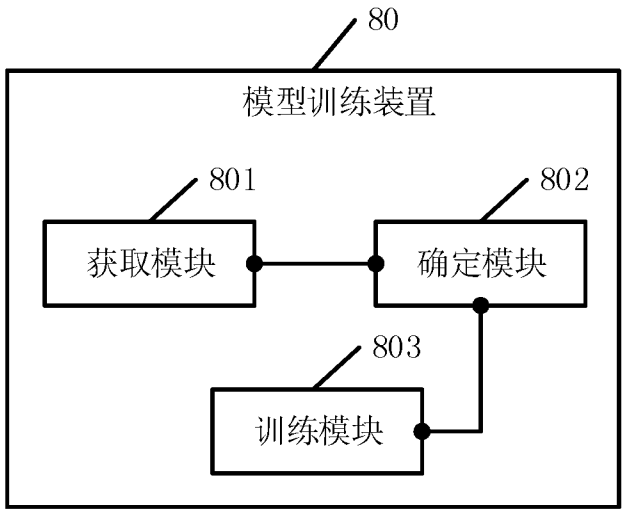
80
模型训练装置
801
获取模块
802
确定模块
803
训练模块

图 8

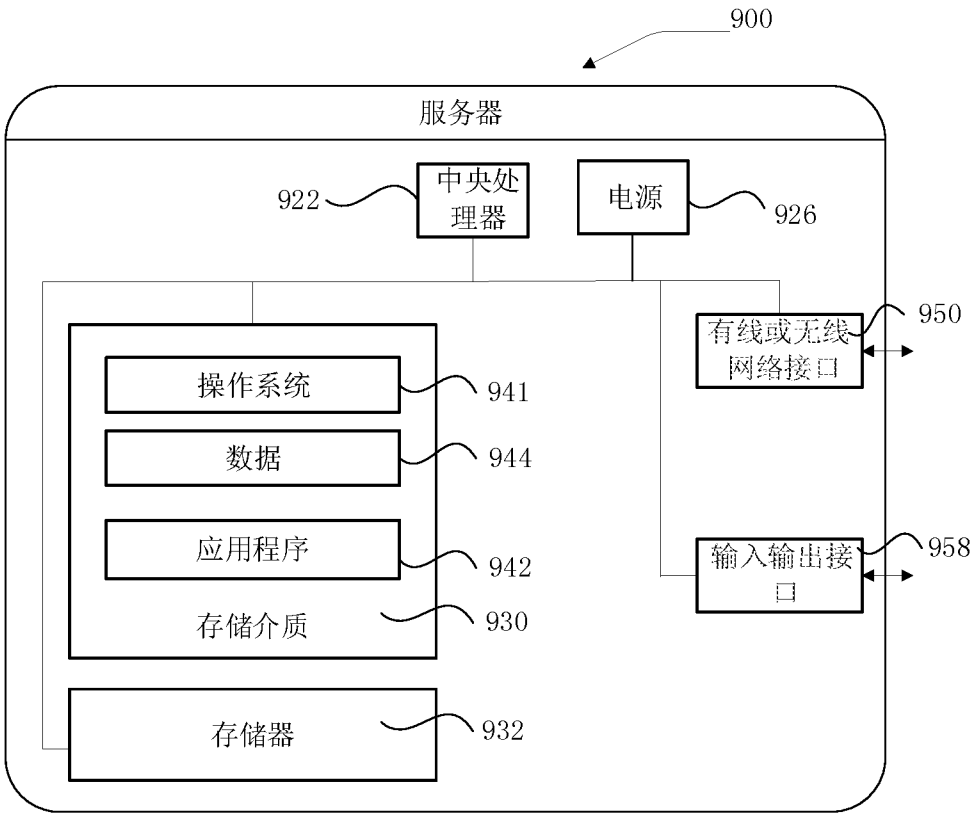
900
服务器
922
中央处理器
电源
926
950
有线或无线网络接口
操作系统
941
数据
944
应用程序
942
存储介质
930
输入输出接口
958
存储器
932

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/077481

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/00(2006.01)i; G06K 9/62(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; CNKI; SIPOABS; DWPI; USTXT; WOTXT; EPTXT: 视频, 对象, 序列, 文本, 字幕, 特征, 匹配, 球员, 合集, 集锦, video, object, sequence, text, subtitles, feature, character, matching, player, collection of the best

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109919078 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 21 June 2019 (2019-06-21) claims 1-15, and description, paragraphs [0005]-[0405]	1-13
X	CN 102427507 A (BEIHANG UNIVERSITY) 25 April 2012 (2012-04-25) description, paragraphs [0008]-[0031]	1-13
X	CN 108229285 A (BEIJING SENSETIME SCIENCE TECHNOLOGY LTD.) 29 June 2018 (2018-06-29) description, paragraphs [0060]-[0224]	1-13
A	CN 102740127 A (SONY CORPORATION) 17 October 2012 (2012-10-17) entire document	1-13

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 April 2020	**29 April 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2020/077481

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109919078 A	21 June 2019	None	
CN 102427507 A	25 April 2012	CN 102427507 B	05 March 2014
CN 108229285 A	29 June 2018	None	
CN 102740127 A	17 October 2012	GB 2489675 A	10 October 2012
		CN 102740127 B	14 December 2016
		US 8745258 B2	03 June 2014
		US 2012254369 A1	04 October 2012

国际检索报告

国际申请号
PCT/CN2020/077481

A. 主题的分类

G06K 9/00(2006.01)i; G06K 9/62(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06K

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNABS;CNTXT;CNKI;SIPOABS;DWPI;USTXT;WOTXT;EPTXT：视频，对象，序列，文本，字幕，特征，匹配，球员，合集，集锦，video，object，sequence，text，subtitles，feature，character，matching，player，collection of the best

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 109919078 A (腾讯科技深圳有限公司) 2019年 6月 21日 (2019-06-21) 权利要求第1-15项，说明书第[0005]-[0405]段	1-13
X	CN 102427507 A (北京航空航天大学) 2012年 4月 25日 (2012-04-25) 说明书第[0008]-[0031]段	1-13
X	CN 108229285 A (北京市商汤科技开发有限公司) 2018年 6月 29日 (2018-06-29) 说明书第[0060]-[0224]段	1-13
A	CN 102740127 A (索尼公司) 2012年 10月 17日 (2012-10-17) 全文	1-13

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 4月 17日	2020年 4月 29日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	李娇
传真号 (86-10)62019451	电话号码 （86-512）88995726

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/077481

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	109919078	A	2019年 6月 21日	无			
CN	102427507	A	2012年 4月 25日	CN	102427507	B	2014年 3月 5日
CN	108229285	A	2018年 6月 29日	无			
CN	102740127	A	2012年 10月 17日	GB	2489675	A	2012年 10月 10日
				CN	102740127	B	2016年 12月 14日
				US	8745258	B2	2014年 6月 3日
				US	2012254369	A1	2012年 10月 4日