

(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
2019 年 1 月 10 日 (10.01.2019)



(10) 国际公布号
WO 2019/007041 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2018/074408

(22) 国际申请日: 2018 年 1 月 29 日 (29.01.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

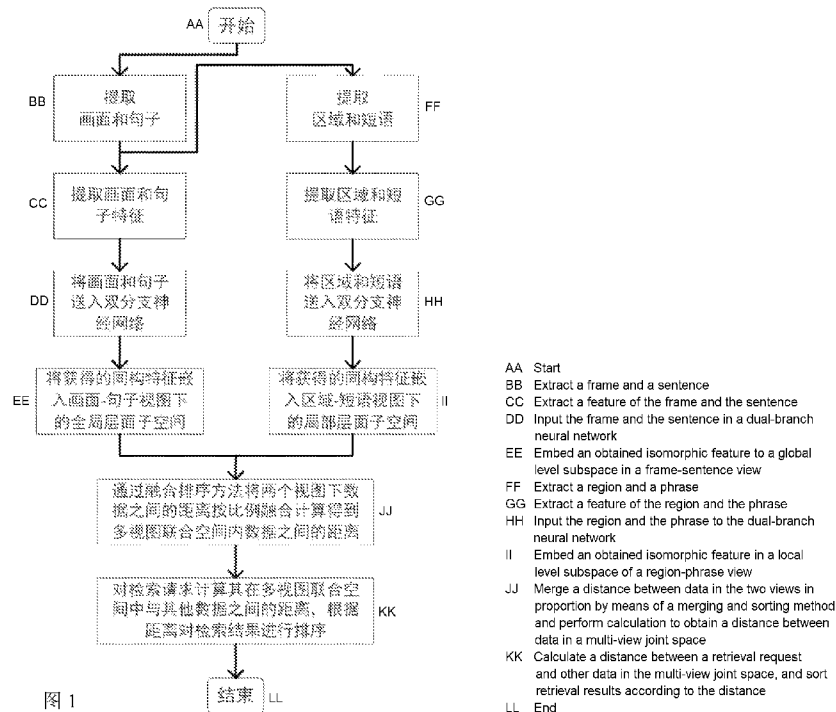
(30) 优先权:
201710545632.X 2017 年 7 月 6 日 (06.07.2017) CN

(71) 申请人: 北京大学深圳研究生院(PEKING UNIVERSITY SHENZHEN GRADUATE SCHOOL) [CN/CN]; 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。

(72) 发明人: 王文敏(WANG, Wenmin); 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。 冉璐(RAN, Lu); 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。 王荣刚(WANG, Ronggang); 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。 李革(LI, Ge); 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。 董胜富(DONG, Shengfu); 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。 王振宇(WANG, Zhenyu); 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。 李英(LI, Ying); 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区,

(54) Title: BIDIRECTIONAL IMAGE-TEXT RETRIEVAL METHOD BASED ON MULTI-VIEW JOINT EMBEDDING SPACE

(54) 发明名称: 基于多视图联合嵌入空间的图像-文本双向检索方法



(57) Abstract: The invention discloses a bidirectional image-text retrieval method based on a multi-view joint embedding space, comprising: performing retrieval with reference to a semantic association relationship between a global level and a local level; obtaining the semantic association relationship between the global level and the local level under a frame-sentence view and a region-phrase view, respectively, and then obtaining from the frame-sentence view semantic association information in a frame and sentence global level subspace; obtaining from the region-phrase view semantic association information in a region and phrase local level subspace;

Guangdong 518055 (CN)。赵辉(ZHAO, Hui); 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。高文(GAO, Wen); 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。

(74) 代理人: 北京万象新悦知识产权代理事务所(普通合伙)(BEIJING WANXIANGXINYUE INTELLECTUAL PROPERTY OFFICE); 中国北京市海淀区颐和园路1号北京资源燕园宾馆六层1625号, Beijing 100080 (CN)。

(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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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))。

processing data by a dual-branch neural network in the two views to obtain an isomorphic feature and embedding the same in a common space, and using a constraint condition to reserve an original semantic relationship of the data during training; and merging two semantic association relationships by means of a multi-view merging and sorting method to obtain an accurate semantic similarity between data, so as to obtain a high precision retrieval result.

(57) 摘要: 本发明公布了一种基于多视图联合嵌入空间的图像-文本双向检索方法, 通过结合全局层面和局部层面的语义关联关系进行检索: 先从画面-句子视图和区域-短语视图下分别获得全局和局部层面的语义关联关系, 在画面-句子视图中获取画面和句子全局层面子空间中的语义关联信息; 在区域-短语视图中获取区域和短语局部层面子空间中的语义关联信息; 两个视图中均通过双分支的神经网络处理数据得到同构特征嵌入共同空间, 在训练中使用约束条件保留数据原有的语义关系; 再通过多视图融合排序方法融合两种语义关联关系得到数据之间更精准的语义相似度, 使得检索结果准确度更高。

基于多视图联合嵌入空间的图像-文本双向检索方法

技术领域

本发明涉及计算机视觉领域，尤其涉及一种基于多视图联合嵌入空间的图像-文本双向检索方法，通过学习得到多视图联合嵌入空间，实现双向图像-文本检索任务。该方法利用不同视图观察数据，获得数据在不同粒度下的语义关联关系，并通过融合排序方法进行融合，获得更精确的语义关联关系，使得检索结果更加准确。

背景技术

随着计算机视觉领域研究的不断发展，出现了大批与图像-文本相关的任务，如图片说明（imagecaption），图片描述（densecaption），以及视觉问题回答（visualquestionanswering）等。所有这些任务都要求计算机充分理解图像和文本的语义信息，并通过学习能够将一种模态数据的语义信息翻译到另一个模态，因此，该类任务的核心问题是如何弥合两种不同模态在语义层面上的差距，找到沟通两种模态数据的方法。不同模态的数据存在于异构的空间中，因而想要直接通过计算数据之间的距离来度量它们的语义相似度是不现实的。为了解决这一问题，目前的主流方法是对两种数据进行变形得到同构的特征，从而将数据嵌入到一个共同的空间，这样它们可以直接被比较。针对使用何种特征表示数据和使用何种方法将数据嵌入到共同空间中，研究者们做了大量的工作。

考虑到一些由深度学习得到的特征在许多计算机视觉领域的任务上取得了很好的成绩，大量的研究者使用这些特征来表示图像或文本，并将它们转化成同构形式从而可以将数据映射到共同空间来完成图像-文本检索任务。然而，这些学习得到的特征仅仅提供了数据全局层面的信息而缺少局部信息的描述，因此只用这些特征来表示数据是无法挖掘更细粒度的数据之间的关联性的，比如图像的区域和文本中的短语之间的关联关系。另一类方法是将图像和文本都切分成小的部分，将它们投影到一个共同的空间，以便于从细粒度的数据中捕捉到局部语义关联信息。

尽管上述方法在某些方面已经取得了很好的表现，但仅仅从单一的视角，即局部视角或全局视角，观察语义关联是无法获得两种模态数据之间全面完整的关联关系的。亦即，同时获得上述两种视角观察得到的语义关联并加以合理利用有利于综合理解数据的语义信息并获得精确的检索结果。但目前尚无能同时获得不同视角的异构数据关联关系，并对这些关系加以融合得到数据之间最终的语义相似度的方法。

发明内容

为了克服上述现有技术的不足，本发明提供一种基于多视图联合嵌入空间的图像-文本双向检索方法，通过结合全局层面和局部层面的信息来完成图像-文本检索任务，从画面-句子视图和区域-短语视图下观察数据，获得全局层面和局部层面的语义关联关系，然后通过融合这两种语义关联关系得到一个精准的语义理解。

本发明的原理是：现有的图像和文本语义信息的获取和理解的方法在某些方面已经取得了很好的表现，但是，现有仅仅从单一的视角观察语义关联的方法是无法获得两种模态数据之间全面完整的关联关系的。本发明通过结合全局层面和局部层面的信息来完成图像-文本检索任务，即融合上述两种视角观察得到的语义关联，获得更全面准确的语义关联关系，将有利于综合理解数据的语义信息并获得精确的检索结果。本发明提出基于多视图联合空间学习框架分成三个部分：画面-句子嵌入模块、区域-短语嵌入模块，以及多视图融合模块。在画面-句子视图嵌入模块中，将一帧图片或文本中的一个完整句子视为基本单位，通过已有的预先训练好的模型获得两种模态数据（即图像数据和文本数据）富含语义信息的特征，利用双分支神经网络学习两组特征的同构特征，这些特征向量使得图像和文本数据能够被映射到共同的空间中。与此同时，在对双分支神经网络进行训练的过程中，需要保持数据之间原有的关联关系，即语义相似度，之后尝试将这种关联关系在全局层面的子空间中通过可计算得到的距离保存下来。在区域-短语视图嵌入模块中，将每一幅画面和每一个句子分别切分成区域和短语，然后使用已有的特征提取方法提取这些局部数据包含语义信息的特征，将它们送入到另一个双分支神经网络中，同样获得保持了语义相似度的同构特征和一个嵌入了细粒度数据并可以直接计算距离的局部层面子空间，以便于探索这些细粒度数据之间的关联关系。

在这两个模块中，为了将异构数据嵌入到共同空间中，两个视图中均采用双分支的神经网络，分别用一个分支处理一个模态的数据，使其变成同构的可进行比较的特征。每一分支的神经网络分别由两层全连接层构成。两个视图中的双分支神经网络结构相同，但是针对不同训练数据训练之后，两个神经网络可以分别针对画面-句子数据，和区域-短语数据提取特征并保留不同视图下的语义关联关系。本发明使用数据向量之间的内积来计算距离，并以此表示数据之间的相似度，并且在训练期间，为了保存语义关联信息，使用了一系列的约束条件来保证语义相关的数据在这一共同空间中有着更邻近的位置关系。

在多视图融合模块中，我们通过按比例结合前两个视图中计算得到的距离来计算图像和文本数据之间在多视图联合空间中的距离。这一最终的距离能够更精确的显示出数据之间的

语义相似度，并且可以作为检索任务的排序依据。需要注意的是，在执行检索任务的过程中，可以单独利用一个视图的语义关联关系进行检索，也可以使用多视图融合后的语义关联关系进行检索，在后续的实验结果中，可以说明多视图融合后得到的语义相似度更能准确的反应数据之间的语义关联关系。

本发明提供的技术方案是：

一种基于多视图联合嵌入空间的图像-文本双向检索方法，通过结合全局层面和局部层面的信息进行图像-文本双向检索；针对数据集 $D = \{D_1, D_2, \dots, D_{|D|}\}$ ，数据集中每一个文档 D_i 包括一张图片 I_i 和一段相关的文本 T_i ， $D_i = (I_i, T_i)$ ，每一段文本由多个句子组成，每一个句子都独立地对相匹配的图片进行描述；在画面-句子视图中，设定 f_i 代表训练图像 I_i 的一幅画面， $\{s_{i1}, s_{i2}, \dots, s_{ik}\}$ 代表 T_i 中的句子集合， k 是文本 T_i 中句子的个数；在区域-短语视图中，设定 r_{im} 代表画面 f_i 提取出的第 m 个区域， p_{in} 代表文本 T_i 中的句子提取出的第 n 个短语；本发明方法首先从画面-句子视图和区域-短语视图下观察数据，分别获得全局层面和局部层面的语义关联关系，然后通过融合这两种语义关联关系得到一个精准的语义理解；具体包括如下步骤：

- 1) 分别提取所有的图像的画面和所有文本中的句子，使用已有的 19 层 VGG (Visual Geometry Group 提出的神经网络结构) 模型提取画面数据的 CNN (Convolutional Neural Network) 特征，使用已有的混合高斯-拉普拉斯混合模型 (HGLMM) 提取句子数据的 FV (Fisher vector) 特征；
- 2) 将步骤 1) 得到的两组特征 (CNN 特征和 FV 特征) 分别送入双分支神经网络的两个分支中，经过训练得到画面和句子数据的同构特征，此时画面和句子被映射到全局层面子空间，并获得画面-句子视图下图像和文本数据的语义关联信息；
- 3) 使用已有的 Faster RCNN 模型 (Faster Region-based Convolutional Network) 提取所有画面的区域 RCNN 特征，使用已有的 Stanford CoreNLP 语法分析器提取所有句子的短语依赖关系 (dependency triplet)，保留含有关键信息的区域和短语特征；
- 4) 将步骤 3) 得到的两组特征 (RCNN 特征和短语的依赖关系) 分别送入另一个双分支神经网络的两个分支中，经过训练得到区域和短语数据的同构特征，此时区域和短语被映射到局部层面子空间，得到区域-短语视图下图像和文本数据的语义关联信息；
- 5) 使用融合排序方法，将步骤 2) 和步骤 4) 得到的不同视图下图像和文本数据的语义关联信息融合起来，计算得到多视图下图像和文本数据在多视图联合空间内的距离，该距离用于度量语义相似度，在检索过程中将其作为排序标准；

6) 对检索请求计算得到该检索请求数据在多视图联合空间中与数据集 D 中另一模态数据 (图像或文本) 之间的距离 (即多视图下图像和文本数据在多视图联合空间内的距离), 根据距离对检索结果进行排序;

由此实现基于多视图联合嵌入空间的图像-文本双向检索。

与现有技术相比, 本发明的有益效果是:

本发明提供一种基于多视图联合嵌入空间的图像-文本双向检索方法, 通过学习得到多视图联合嵌入空间, 实现双向图像-文本检索任务。该方法利用不同视图观察数据, 获得数据在不同粒度下的语义关联关系, 并通过融合排序方法进行融合, 获得更精确的语义关联关系, 使得检索结果更加准确。具体地, 本发明使用双分支神经网络分别将图像和文本数据映射到画面-句子视图的全局层面子空间和区域-短语视图的局部层面子空间, 获得全局层面和局部层面的语义关联关系, 根据这两组语义关联关系可以单独完成图像-文本双向检索任务, 但所得到的检索结果并不全面。本发明提出多视图融合排序方法, 将两个视图下的语义关联关系融合在一起, 共同计算数据之间在多视图联合空间中的距离, 得到的数据之间的距离关系能够更精准的反应数据的语义相似度, 使得检索结果的准确度更高。具体地, 本发明具有如下技术优势:

(一) 本发明从多个视图观察不同模态数据之间的高层语义关联关系, 并将这些关联关系融合在一起形成多视图下的语义关联关系, 而现有图像-文本检索方法没有此考虑; 采用本发明方法可以学习到数据在不同粒度下存在的语义信息, 从而能有效的提取更精确的语义信息, 获得准确度更高的检索结果;

(二) 本发明使用融合排序方法能够将不同视图的语义关联关系融合在一起, 使得数据在多视图联合空间中的相对距离能够很好地综合不同视图下的语义关系, 最终获得更精确的语义相似度;

(三) 本发明采用双分支神经网络, 其作用在于不同模态的数据其特征是异构的, 无法直接进行比较或计算距离, 通过双分支神经网络, 每一分支处理一个模态的数据, 将不同模态的数据变形成为同构特征, 使异构数据同时存在于一个共同空间, 可以直接计算距离;

(四) 本发明采用一系列约束条件, 为了获得同构特征, 利用双分支神经网络对数据进行转换, 但是在转换过程中, 数据原有的语义关系需要得到保留; 利用基于间隔的随机损失函数将语义相关数据之间的距离与语义无关数据之间的距离范围拉开间隔, 确保在共同空间

中数据之间的语义相似度信息得到保留。

附图说明

图 1 是本发明中基于多视图联合嵌入空间进行图像-文本双向检索的流程框图。

图 2 是本发明实施例中多视图联合嵌入空间学习过程的示意图；

其中，VGG 是 19 层 VGG 模型，提取画面的 CNN 特征，HGLMM 是混合高斯-拉普拉斯混合模型（Hybrid Gaussian-Laplacian mixture model）提取句子的 Fisher Vector 特征，RCNN 是 Faster RCNN 模型提取区域的 RCNN 特征，Parser 是 Stanford CoreNLP 语法分析器提取短语的依赖三元组。神经网络均由两层全连接层组成，每个视图下的两个神经网络组成双分支神经网络。

图 3 模态间一致性与模态内一致性示意图；

其中，（a）为模态间一致性表示，（b）为模态内一致性表示。

图 4 是本发明实施例提供的采用本发明方法在 Pascal1K 数据集下得到的图像-文本检索结果。

具体实施方式

下面结合附图，通过实施例进一步描述本发明，但不以任何方式限制本发明的范围。

本发明提供一种基于多视图联合嵌入空间的图像-文本双向检索方法，通过结合全局层面和局部层面的语义关联关系进行检索；先从画面-句子视图和区域-短语视图下分别获得全局和局部层面的语义关联关系，在画面-句子视图中，将画面和句子嵌入到全局层面的子空间中获取画面和句子的语义关联信息；在区域-短语视图中，提取画面的每一个区域和句子的每一个短语，嵌入到局部层面的子空间中获取区域和短语的语义关联信息；两个视图中均通过双分支的神经网络处理数据得到同构特征嵌入共同空间，在训练中使用约束条件保留数据原有的语义关系；再通过多视图融合排序方法融合两种语义关联关系得到数据之间更精准的语义相似度。

我们利用数据集 $D = \{D_1, D_2, \dots, D_{|D|}\}$ 对图像-文本检索问题加以描述，在这一数据集中，每一个文档包括一张图片和一段相关的文本，如 $D_i = (I_i, T_i)$ ，每一个文本由若干个句子组成，每一个句子都独立地对相匹配的图片进行描述。在画面-句子视图中，我们用 f_i 代表训练图像 I_i 的一幅画面，用 $\{s_{i1}, s_{i2}, \dots, s_{ik}\}$ 代表 T_i 中的句子集合（ k 是文本 T_i 中句子的个数）。在区域-短语视图中，我们用 r_{im} 代表画面 f_i 提取出的第 m 个区域，用 p_{in} 代表文本 T_i 中的句子提取出的第 n

个短语。接下来我们将详细描述两个视图中使用的约束条件，以及最后多视图融合模块的融合排序方法。

1、画面-句子视图

我们将画面和句子分别送入双分支神经网络中，并且获得在全局层面子空间的同构特征。在训练神经网络的过程中，为了保留模态间一致性和模态内一致性，我们提出了基于间隔的随机损失函数。为了将数据放入画面-句子视图中进行处理，我们做了如下特征提取：

对于图像，我们使用 19 层 VGG 模型提取得到的 4096 维 CNN 特征（向量）作为画面的原始特征；对于文本，我们使用混合高斯-拉普拉斯混合模型（Hybrid Gaussian-Laplacian mixture model, HGLMM）提取得到的 Fisher vector (FV) 特征（向量）作为句子的原始特征，为了计算方便，我们用 PCA（Principal Components Analysis，主成分分析）将最初的 18000 维 FV 特征向量降维至 4999 维。

1) 模态间一致性

对于训练画面 f_i ，所有的句子可以分成两个集合，一个集合包含了所有与训练画面匹配的句子，一个集合包含了所有与训练画面不匹配的句子。我们可以推测出一个合理的一致性要求，即在画面-句子视图中，画面 f_i 和在匹配集合中的句子之间的距离必须比画面 f_i 和在不匹配集合中的句子之间的距离小，且距离的差距需要大于间隔 m 。具体数学表示如式（1）：

$$d(f_i, s_{ix}) + m < d(f_i, s_{jy}) \text{ if } i \neq j \quad (1)$$

式（1）中， $d(f_i, s_{ix})$ 表示画面 f_i 与在其匹配集合中的句子 s_{ix} 之间的距离； $d(f_i, s_{jy})$ 表示画面 f_i 与在其不匹配集合中的句子 s_{jy} 之间的距离。

类似的约束可以应用在训练句子 s_{ix} 上：

$$d(f_i, s_{ix}) + m < d(f_j, s_{ix}) \text{ if } i \neq j \quad (2)$$

式（2）中， $d(f_i, s_{ix})$ 表示句子 s_{ix} 与在其匹配集合中的画面 f_i 之间的距离； $d(f_j, s_{ix})$ 表示句子 s_{ix} 与在其不匹配集合中的画面 f_j 之间的距离。

2) 模态内一致性

在训练过程中，除了考虑到数据的模态间一致性，我们还需要针对数据集中同一画面伴随的若干个句子做一些约束，我们称之为模态内一致性。具体来说，对于共享相同含义即描述相同画面的句子而言，他们需要紧密的联系在一起，并且能够与其他的句子区分开来。

为了实现模态内一致性，我们使用了如下约束：

$$d(s_{ix}, s_{iy}) + m < d(s_{ix}, s_{jz}) \text{ if } i \neq j \quad (3)$$

式 (3) 中, $d(s_{ix}, s_{iy})$ 表示描述同一个画面 f_i 的句子 s_{ix} 与 s_{iy} 之间的距离; $d(s_{ix}, s_{jz})$ 表示描述画面 f_i 的句子 s_{ix} 与描述画面 f_j 的句子 s_{jz} 之间的距离。

尽管我们应该对画面作出类似公式 (3) 的约束, 即描述同一个句子的画面之间的距离应该更近, 但目前我们使用的数据集中, 还难以确定若干画面是否描述了相同的句子, 因此我们不采用这一项约束, 而只对描述同一个画面的句子之间的距离进行公式 (3) 的约束。

结合上述的约束条件, 我们最终总结出了在画面-句子视图上的损失函数:

$$\begin{aligned} \psi_{frame-sentence} &= \sum_{i,j,x,y} \max[0, m + d(f_i, s_{ix}) - d(f_i, s_{jy})] \\ &+ \lambda_1 \sum_{i,j,x} \max[0, m + d(f_i, s_{ix}) - d(f_j, s_{ix})] \\ &+ \lambda_2 \sum_{i,j,x,y,z} \max[0, m + d(s_{ix}, s_{iy}) - d(s_{ix}, s_{jz})] \end{aligned} \quad (4)$$

这里, 间隔 m 可以根据所采用的距离作出调整, 为了便于优化, 在这里我们将其固定为 $m = 0.1$ 并将它应用在所有的训练样本中, 与此同时, 通过实验我们发现当 $\lambda_1 = 2$ 且 $\lambda_2 = 0.2$ 时我们能够获得最好的实验结果。

2、区域-短语视图

在这一视图中, 我们希望能够挖掘存在于区域和短语之间的细粒度语义关联关系。我们通过已有的模型提取得到区域和短语特征, 对于区域, 我们提取画面中得分最高的 19 个区域的 4096 维 RCNN 特征; 对于短语, 我们利用语法分析器得到依赖树结构, 并选择包含关键语义信息的短语。我们用 1-of-k 编码向量 $\mathbf{w}$ 来代表每一个单词, 并将 p_{jy} 用依赖关系三元组 $(R, \mathbf{w}_1, \mathbf{w}_2)$ 表示的短语映射到嵌入空间中, 如式 (5):

$$p_{jy} = f\left(W_R \begin{bmatrix} W_e \mathbf{w}_1 \\ W_e \mathbf{w}_2 \end{bmatrix} + b_R\right) \quad (5)$$

式 (5) 中, W_e 是一个 $400000 \times d$ 的矩阵, 用于将 1-of-k 向量编码成一个 d 维的单词向量, 其中 400000 是字典的单词数目, 这里我们设定 $d=200$ 。注意每一个关系 R 都有单独的权重 W_R 和偏移量 b_R , 并且每一个句子提取的短语数是不一样的。

在这个视图中，我们利用双分支神经网络将图像和文本数据射到区域-短语视图的局部子空间中。在训练网络的过程中，我们要求在匹配的图像文本对中的区域和短语之间的距离要比在不匹配对中的区域和短语之间的距离小。在计算这一视图下的数据映射到局部层面子空间的损失函数表示如下：

$$\psi_{region-phrase} = \sum_{i,j,x,y} \kappa_{ij} \max[0, 1 - \eta_{ij} \times d(r_{ix}, p_{jy})] \quad (6)$$

式（6）中， $d(r_{ix}, p_{jy})$ 表示区域 r_{ix} 和短语 p_{jy} 之间的距离，我们定义 η_{ij} 在 $i = j$ 的时候等于+1，在 $i \neq j$ 时等于-1，常量 κ_{ij} 用于根据 η_{ij} 的正负个数进行归一化。

3、多视图融合模块

在画面-句子嵌入模块和区域-短语嵌入模块分别通过学习得到各自的嵌入空间之后，我们可以借助这两个空间中的信息来获得多视图下的数据间距离。为了获得图像 I_i 和文本 T_j 之间更精确的语义相似度，我们按比例结合前两个视图下计算得到的距离作为多视图联合空间中两个数据之间的最后距离：

$$d_{multi-view}(I_i, T_j) = d_{frame-sentence}(I_i, T_j) + \lambda d_{region-phrase}(I_i, T_j) \quad (7)$$

式（7）中， $d_{frame-sentence}(I_i, T_j)$ 表示图像 I_i 和文本 T_j 之间在画面-句子视图中的距离， $d_{region-phrase}(I_i, T_j)$ 表示图像 I_i 和文本 T_j 之间在区域-短语视图中的距离， $d_{multi-view}(I_i, T_j)$ 表示图像 I_i 和文本 T_j 之间在多视图融合后的距离。权重 λ 用于平衡画面-句子视图和区域-短语视图距离的比例，经过实验，我们发现 $\lambda = 0.6$ 能够产生很好的效果。

图 1 为多视图联合嵌入空间学习框架流程图。

图 2 展示了本发明提出的学习框架示意图。该框架分三个部分，画面-句子视图中，我们将画面和句子嵌入到全局层面的子空间中。在区域-短语视图中，我们提取画面和句子中小的成分，将这些成分嵌入到局部层面的子空间中。每一个视图中，我们都用双分支的神经网络处理数据以使它们变成同构特征嵌入共同空间。在多视图融合模块，提出了多视图融合排序方法来融合两个视图分析得到的语义关联信息得到数据之间最终的距离关系。

图 3 展示了在画面-句子视图下模态间一致性（左）和模态内一致性（右）的示意图。正方形代表画面，圆形代表句子，相同颜色的图像表示的是同一个语义信息。模态间一致性含义为在画面-句子视图中，一个画面（黑色正方形）和与其匹配的句子（黑色圆形）之间的距离必须比该画面（黑色正方形）和与其不匹配的句子（灰色圆形）之间的距离小，此约束对句子同样适用。模态内一致性含义为在画面-句子视图中，一个句子（黑色正方形）和与其语

义相近的句子（其他黑色正方形）之间的距离必须比该句子（黑色正方形）和与其语义无关的句子（灰色正方形）之间的距离小。

图 4 展示了一个图像-文本检索的实际案例，分别给出了画面-句子视图，区域-短语视图和多视图在针对左上角图像返回的前五个句子的检索结果，正确的检索结果用加粗表示。在这个例子中，我们可以看出，画面-句子视图仅仅能检索到全局层面理解后的句子，但是由于它不能辨别区域中的内容，所以会把正确匹配的句子和一些有相似含义但并不包括不正确个体的句子混淆。对于区域-短语视图，它返回了一些包含有正确的个体但个体之间关系不准确的句子。比如这一视图下的第三个句子，它可以辨认出‘a young girl’，但是却误解了女孩和自行车之间的关系是‘riding’，最终返回了一个错误的句子。但是在融合后的多视图中，它可以同时捕捉到全局层面的语义关联关系和局部层面的语义关联关系，因而多视图的检索效果是准确度最高的。

表 1. Pascal1K 数据集下实施例图像-文本双向检索结果

方法	Pascal1K					
	图像-文本检索			文本-图像检索		
	R@1	R@5	R@10	R@1	R@5	R@10
随机排列	4.0	9.0	12.0	1.6	5.2	10.6
SDT-RNN	23.0	45.0	63.0	16.4	46.6	65.6
kCCA	21.0	47.0	61.0	16.4	41.4	58.0
DeViSE	17.0	57.0	68.0	21.0	54.6	72.4
DCCA	-	-	-	-	-	-
VQA-A	-	-	-	-	-	-
DFE	39.0	68.0	79.0	23.6	65.2	79.8
本发明（仅画面-句子视图）	30.0	62.0	76.0	21.8	65.2	83.2
本发明（仅区域-短语视图）	35.0	67.0	86.0	29.4	67.8	82.6
本发明（多视图）	47.0	83.0	93.0	29.0	73.4	87.0

表 2. Flickr8K 数据集下实施例图像-文本双向检索结果

方法	Flickr8K					
	图像-文本检索			文本-图像检索		
	R@1	R@5	R@10	R@1	R@5	R@10
随机排列	0.1	0.6	1.1	0.1	0.5	1.0
SDT-RNN	4.5	18.0	28.6	6.1	18.5	29.0
kCCA	-	-	-	-	-	-

DeViSE	4.8	16.5	27.3	5.9	20.1	29.6
DCCA	28.2	56.1	69.8	26.3	54.0	67.5
VQA-A	24.3	52.2	65.2	17.2	42.8	57.2
DFE	12.6	32.9	44.0	9.7	29.6	42.5
本发明（仅画面-句子视图）	20.7	43.7	57.4	15.8	32.7	47.4
本发明（仅区域-短语视图）	17.6	39.8	49.0	12.5	30.1	48.6
本发明（多视图）	24.5	60.2	73.5	18.9	48.9	68.2

表 1 和表 2 分别给出了本发明在 Pascal1K 和 Flickr8K 上通过图像-文本检索和文本-图像检索进行实验验证的结果。为了评价检索效果，我们遵循了标准的排序度量标准，使用 Recall@K，即正确匹配的数据排在前K(K = 1,5,10)个检索结果中的概率，来对检索准确性进行度量。图中列出了本发明与其他现有先进算法的效果比较，包括 SDT-RNN（Semantic Dependency Trees -Recurrent Neural Networks，语义依赖树-循环神经网络），kCCA（kernel Canonical Correlation Analysis，核典型相关分析），DeViSE（Deep Visual-Semantic Embedding，深度视觉语义嵌入），DCCA（Deep Canonical Correlation Analysis，深度典型相关分析），VQA-A（VisualQuestion Answering-agnostic，视觉问题回答不可知论），DFE（Deep Fragment Embedding，深度片段嵌入）。

从表 1 和表 2 我们可以看出，本发明方法和其他对比方法相比较而言效果更好。此外，我们还分别展示了两种单独视图的检索效果和多视图融合后的检索效果，从数据中可以看出，结合两个视图之后的多视图融合方法检索效果更好。这一结果证明了单独的两个视图之间彼此是互补的关系，所以将二者结合起来之后可以获得更精准全面的语义关联，检索效果会更好。

需要注意的是，公布实施例的目的在于帮助进一步理解本发明，但是本领域的技术人员可以理解：在不脱离本发明及所附权利要求的精神和范围内，各种替换和修改都是可能的。因此，本发明不应局限于实施例所公开的内容，本发明要求保护的范围以权利要求书界定的范围为准。

权 利 要 求 书

1、一种基于多视图联合嵌入空间的图像-文本双向检索方法，通过结合全局层面和局部层面的语义关联关系进行图像-文本双向检索；针对数据集 $D = \{D_1, D_2, \dots, D_{|D|}\}$ ，数据集中每一个文档 D_i 包括一张图片 I_i 和一段相关的文本 T_i ，表示为 $D_i = (I_i, T_i)$ ，每一段文本由多个句子组成，每一个句子均独立描述相匹配的图片；在画面-句子视图中，设定 f_i 代表训练图像 I_i 的一幅画面， $\{s_{i1}, s_{i2}, \dots, s_{ik}\}$ 代表 T_i 中的句子集合， k 是文本 T_i 中句子的个数；在区域-短语视图中，设定 r_{im} 代表画面 f_i 提取出的第 m 个区域， p_{in} 代表文本 T_i 中的句子提取出的第 n 个短语；所述双向检索方法首先从画面-句子视图和区域-短语视图下分别获得全局层面和局部层面的语义关联关系，然后通过融合两种语义关联关系得到一个精准的语义理解；具体包括如下步骤：

- 1) 分别提取所有图像的画面和所有文本中的句子，分别送入模型中提取数据的特征，得到画面的 CNN 特征和句子的 FV 特征；
- 2) 将步骤 1) 得到的画面的 CNN 特征和句子的 FV 特征分别送入双分支神经网络的两个分支中，经过训练得到画面和句子数据的同构特征，此时画面和句子被映射到全局层面子空间，并获得画面-句子视图下图像和文本数据的语义关联信息；
- 3) 使用 RCNN 模型和语法分析器分别提取所有画面的区域 RCNN 特征和所有句子的短语依赖关系，保留含有关键信息的区域和短语的特征；
- 4) 将步骤 3) 得到的区域和短语的特征分别送入另一个双分支神经网络的两个分支中，经过训练得到区域和短语数据的同构特征，此时区域和短语被映射到局部层面子空间，得到区域-短语视图下图像和文本数据的语义关联信息；
- 5) 使用融合排序方法，将步骤 2) 和步骤 4) 得到的不同视图下图像和文本数据的语义关联信息进行融合，计算得到多视图下图像和文本数据在多视图联合空间内的距离，该距离用于度量语义相似度，在检索过程中作为排序标准；
- 6) 对检索请求计算得到该检索请求数据在多视图联合空间中与数据集中另一模态数据之间的距离，根据距离对检索结果进行排序；

由此实现基于多视图联合嵌入空间的图像-文本双向检索。

2、如权利要求 1 所述图像-文本双向检索方法，其特征是，步骤 1) 提取特征具体是：对于图像，使用 19 层 VGG 模型提取得到 4096 维 CNN 特征向量，作为画面的原始特征；对于文本，使用混合高斯-拉普拉斯混合模型提取得到 FV 特征向量，作为句子的原始特征；并通

过 PCA 将特征向量由 18000 维降维至 4999 维。

3、如权利要求 1 所述图像-文本双向检索方法，其特征是，步骤 2) 将特征分别送入双分支神经网络的两个分支中进行训练，得到画面和句子数据的同构特征，训练过程中设定约束条件以保留模态间一致性和模态内一致性，采用基于间隔的随机损失函数；具体包括：

A. 训练画面 f_i ：将所有句子分成匹配集合和不匹配集合，匹配集合包含所有与训练画面匹配的句子，不匹配集合包含所有与训练画面不匹配的句子；设定一致性约束要求为：在画面-句子视图中，画面 f_i 和在匹配集合中的句子之间的距离必须比画面 f_i 和在不匹配集合中的句子之间的距离小，且距离的差距需要大于间隔 m ，表示如式 (1)：

$$d(f_i, s_{ix}) + m < d(f_i, s_{jy}) \text{ if } i \neq j \quad (1)$$

式 (1) 中， $d(f_i, s_{ix})$ 表示画面 f_i 与在其匹配集合中的句子 s_{ix} 之间的距离； $d(f_i, s_{jy})$ 表示画面 f_i 与在其不匹配集合中的句子 s_{jy} 之间的距离；

B. 将式 (2) 的约束应用在训练句子 s_{ix} 上：

$$d(f_i, s_{ix}) + m < d(f_j, s_{ix}) \text{ if } i \neq j \quad (2)$$

式 (2) 中， $d(f_i, s_{ix})$ 表示句子 s_{ix} 与在其匹配集合中的画面 f_i 之间的距离； $d(f_j, s_{ix})$ 表示句子 s_{ix} 与在其不匹配集合中的画面 f_j 之间的距离；

C. 针对数据集中同一画面伴随的多个句子设定约束条件，表示为式 (3)：

$$d(s_{ix}, s_{iy}) + m < d(s_{ix}, s_{jz}) \text{ if } i \neq j \quad (3)$$

式 (3) 中， $d(s_{ix}, s_{iy})$ 表示描述同一个画面 f_i 的句子 s_{ix} 与 s_{iy} 之间的距离； $d(s_{ix}, s_{jz})$ 表示描述画面 f_i 的句子 s_{ix} 与描述画面 f_j 的句子 s_{jz} 之间的距离；

D. 建立在画面-句子视图上的损失函数如式 (4)：

$$\begin{aligned}
& \psi_{\text{frame-sentence}} \\
&= \sum_{i,j,x,y} \max[0, m + d(f_i, s_{ix}) - d(f_i, s_{iy})] \\
&+ \lambda_1 \sum_{i,j,x} \max[0, m + d(f_i, s_{ix}) - d(f_j, s_{ix})] \\
&+ \lambda_2 \sum_{i,j,x,y,z} \max[0, m + d(s_{ix}, s_{iy}) - d(s_{ix}, s_{jz})]
\end{aligned} \tag{4}$$

其中， m 为间隔，可根据所采用的距离作出调整。

4、如权利要求3所述图像-文本双向检索方法，其特征是，间隔 m 取值为0.1；参数 λ_1 取值为2；参数 λ_2 取值为0.2。

5、如权利要求1所述图像-文本双向检索方法，其特征是，步骤3)通过模型提取得到区域和短语特征，具体是：

对于区域，提取画面中得分最高的19个区域的4096维RCNN特征；

对于短语，利用语法分析器得到依赖树结构，选择包含关键语义信息的短语；

用1-of-k编码向量 w 来代表每一个单词，并将 p_{jy} 用依赖关系三元组 (R, w_1, w_2) 表示的短语映射到嵌入空间中，表示为式(5)：

$$p_{jy} = f\left(W_R \begin{bmatrix} W_e w_1 \\ W_e w_2 \end{bmatrix} + b_R\right) \tag{5}$$

式(5)中， W_e 是一个400000× d 的矩阵，用于将1-of-k向量编码成一个 d 维的单词向量，其中400000是字典的单词数目；每一个关系 R 均有单独的权重 W_R 和偏移量 b_R ，并且每一个句子提取的短语数不同。

6、如权利要求1所述图像-文本双向检索方法，其特征是，步骤4)将区域和短语的特征分别送入双分支神经网络的两个分支中进行训练，得到区域和短语数据的同构特征；具体地，在训练过程中，设定条件为：在匹配的图像文本对中的区域和短语之间的距离比在不匹配对中的区域和短语之间的距离小；通过式(6)的损失函数计算区域-短语视图下的数据并映射到局部层面子空间：

$$\psi_{\text{region-phrase}} = \sum_{i,j,x,y} \kappa_{ij} \max[0, 1 - \eta_{ij} \times d(r_{ix}, p_{jy})] \tag{6}$$

式 (6) 中, $d(r_{ix}, p_{jy})$ 表示区域 r_{ix} 和短语 p_{jy} 之间的距离; 设定 η_{ij} 在 $i = j$ 的时候等于 +1, 在 $i \neq j$ 时等于 -1, 常量 κ_{ij} 根据 η_{ij} 的正负个数进行归一化。

7、如权利要求 1 所述图像-文本双向检索方法, 其特征是, 步骤 5) 通过式 (7) 按比例结合两个视图下计算得到的距离, 作为多视图联合空间中两个数据之间的距离:

$$d_{\text{multi-view}}(I_i, T_j) = d_{\text{frame-sentence}}(I_i, T_j) + \lambda d_{\text{region-phrase}}(I_i, T_j) \quad (7)$$

式 (7) 中, $d_{\text{frame-sentence}}(I_i, T_j)$ 表示图像 I_i 和文本 T_j 之间在画面-句子视图中的距离; $d_{\text{region-phrase}}(I_i, T_j)$ 表示图像 I_i 和文本 T_j 之间在区域-短语视图中的距离; $d_{\text{multi-view}}(I_i, T_j)$ 表示图像 I_i 和文本 T_j 之间在多视图融合后的距离; 权重 λ 用于平衡画面-句子视图和区域-短语视图距离的比例。

8、如权利要求 7 所述图像-文本双向检索方法, 其特征是, 权重 λ 取值为 0.6。

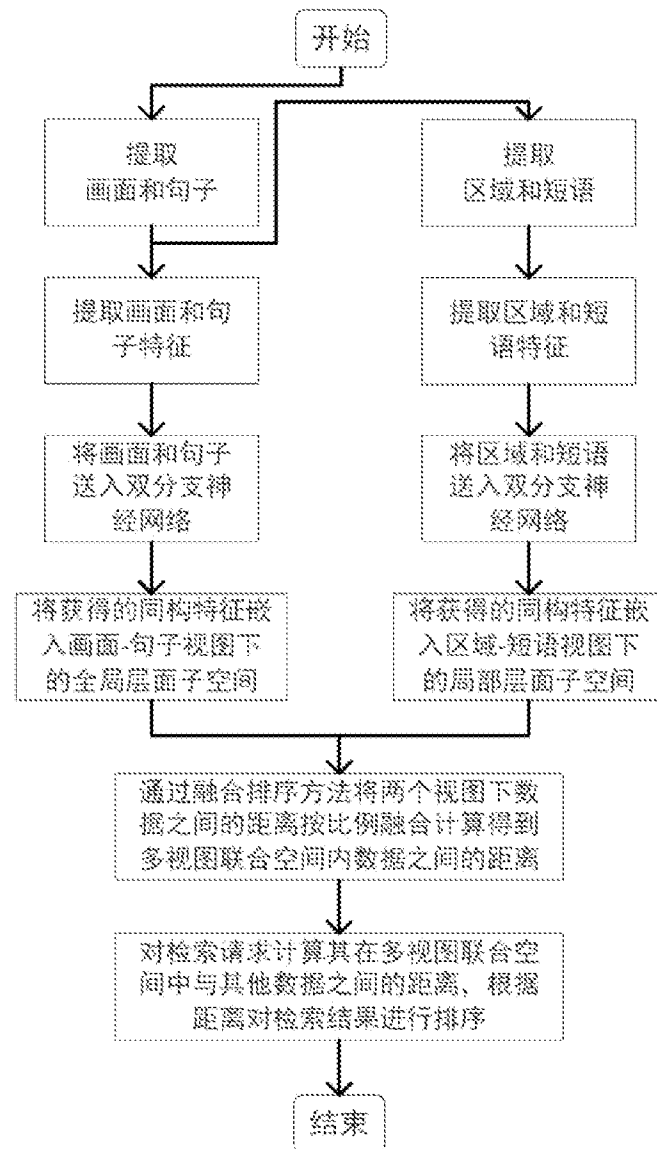


图 1

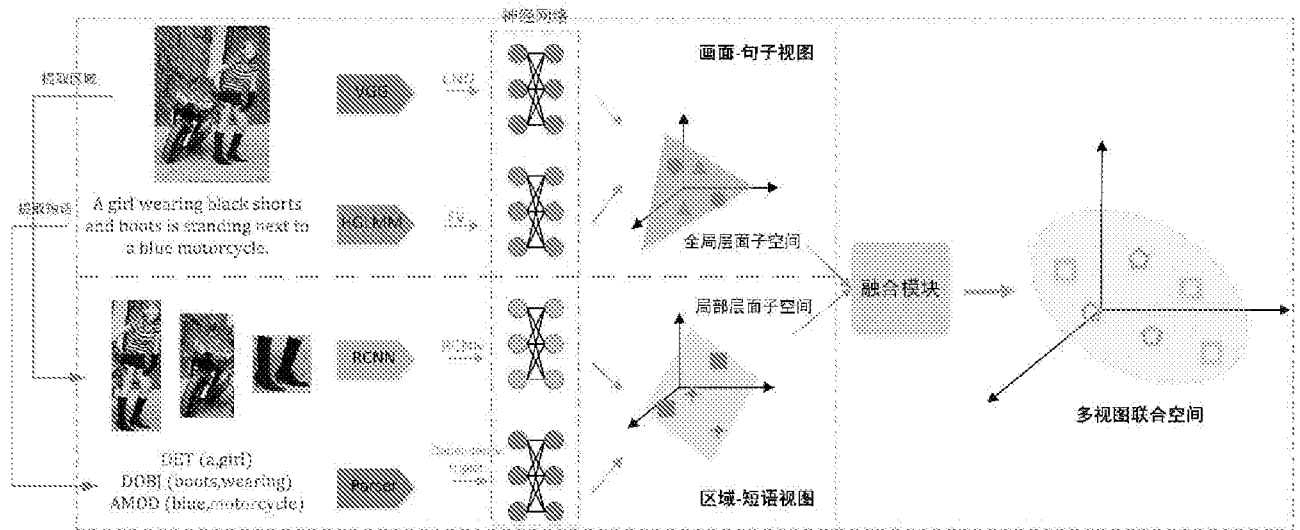


图 2

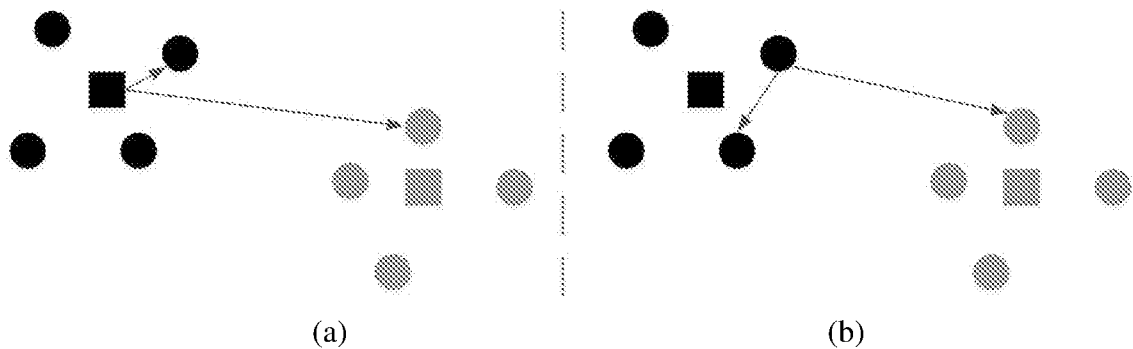
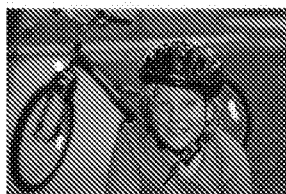


图 3



多视图:

1. A young blonde girl wearing a smile and a bicycle helmet.
2. A smiling young blonde girl wearing a helmet sits in front of her bicycle.
3. A young girl wearing a helmet riding a pony.
4. A little girl with a bike helmet posing by her bicycle.
5. A young blonde girl wearing a helmet sits with a red bicycle behind her.

画面-句子视图:

1. A young girl wearing a bike helmet with a bicycle in the background.
2. A motor-cross rider trying to catch his bike.
3. A man and his bicycle laying down on the green grass.
4. A man wearing a helmet does a wheelie on a motorcycle as a crowd watches.
5. A woman rides as a bicycle drives pulls her in traffic.

区域-短语视图:

1. A smiling young blonde girl wearing a helmet sits in front of her bicycle.
2. A young blonde girl wearing a smile and a bicycle helmet.
3. A young girl wearing a helmet riding a pony.
4. Small boy and puppy standing in front of a film poster.
5. A girl with arms crossed sitting on a sofa.

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/074408

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (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)

G06F, 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; DWPI; SIPOABS; IEEE; USTXT: 检索, 搜索, 文本, 距离, 视图, 句子, 画, 图, 语义, retrieve, search, text, distance, view, sentence, image, graphic, semantic

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106202413 A (PEKING UNIVERSITY SHENZHEN GRADUATE SCHOOL), 07 December 2016 (07.12.2016), entire document	1-8
A	CN 102138140 A (DOSSIERVER INC.), 27 July 2011 (27.07.2011), entire document	1-8
A	US 7814040 B1 (UNIV NEW YORK STATE RES FOUND), 12 October 2010 (12.10.2010), entire document	1-8
PX	CN 107330100 A (PEKING UNIVERSITY SHENZHEN GRADUATE SCHOOL), 07 November 2017 (07.11.2017), claims 1-8	1-8

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 23 April 2018	Date of mailing of the international search report 02 May 2018
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer YANG Jie Telephone No. 86-010-62411825

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/074408

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 106202413 A	07 December 2016	WO 2018010365 A1	18 January 2018
CN 102138140 A	27 July 2011	EP 2307983 A1	13 April 2011
		CA 2729716 A1	07 January 2010
		EP 2307983 A4	14 November 2012
		WO 2010000064 A1	07 January 2010
		US 2010005061 A1	07 January 2010
		US 8060513 B2	15 November 2011
US 7814040 B1	12 October 2010	US 8204842 B1	19 June 2012
		US 9280562 B1	08 March 2016
CN 107330100 A	07 November 2017	None	

国际检索报告

国际申请号

PCT/CN2018/074408

A. 主题的分类

G06F 17/30(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)

G06F, 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;DWPI;SIPOABS;IEEE;USTXT:检索, 搜索, 文本, 距离, 视图, 句子, 画, 图, 语义, retrieve, search, text, distance, view, sentence, image, graphic, semantic

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
A	CN 106202413 A (北京大学深圳研究生院) 2016年 12月 7日 (2016 - 12 - 07) 全文	1-8
A	CN 102138140 A (多斯维公司) 2011年 7月 27日 (2011 - 07 - 27) 全文	1-8
A	US 7814040 B1 (UNIV NEW YORK STATE RES FOUND) 2010年 10月 12日 (2010 - 10 - 12) 全文	1-8
PX	CN 107330100 A (北京大学深圳研究生院) 2017年 11月 7日 (2017 - 11 - 07) 权利要求1-8	1-8

☐ 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

2018年 4月 23日

国际检索报告邮寄日期

2018年 5月 2日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

杨洁

电话号码 86-010-62411825

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/074408

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	106202413	A	2016年 12月 7日	WO	2018010365	A1	2018年 1月 18日
CN	102138140	A	2011年 7月 27日	EP	2307983	A1	2011年 4月 13日
				CA	2729716	A1	2010年 1月 7日
				EP	2307983	A4	2012年 11月 14日
				WO	2010000064	A1	2010年 1月 7日
				US	2010005061	A1	2010年 1月 7日
				US	8060513	B2	2011年 11月 15日
US	7814040	B1	2010年 10月 12日	US	8204842	B1	2012年 6月 19日
				US	9280562	B1	2016年 3月 8日
CN	107330100	A	2017年 11月 7日	无			

表 PCT/ISA/210 (同族专利附件) (2015年1月)