

(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 1 月 2 日 (02.01.2020)



(10) 国际公布号
WO 2020/001048 A1

(51) 国际专利分类号:
G06F 16/48 (2019.01)

(21) 国际申请号: PCT/CN2019/076400

(22) 国际申请日: 2019 年 2 月 28 日 (28.02.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810695406.4 2018年6月29日 (29.06.2018) CN
201811106799.7 2018年9月21日 (21.09.2018) 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)。 夏雅娴(XIA, Yaxian); 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)。 韩梁(HAN, Liang); 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)。 高文(GAO,

(54) Title: DOUBLE SEMANTIC SPACE-BASED ADVERSARIAL CROSS-MEDIA RETRIEVAL METHOD

(54) 发明名称: 一种基于双语义空间的对抗性跨媒体检索方法

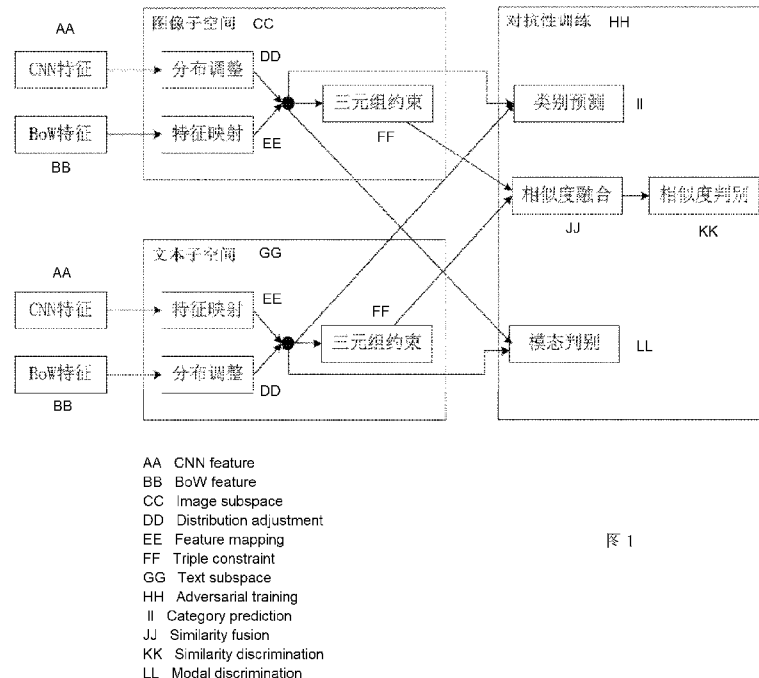


图 1

(57) Abstract: Disclosed is a double semantic space-based adversarial cross-media retrieval method, which relates to the technical fields of pattern recognition, natural language processing, multimedia retrieval and the like, and comprises: a feature generation process, a double semantic space construction process, and an adversarial semantic space optimization process. According to the present invention, by establishing an isomorphic double semantic space, namely a text subspace and an image subspace, a semantic gap is eliminated while the original image and text information are retained to the maximum extent; and the isomorphic subspace data distribution is optimized

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有限公司(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))。

by means of adversarial training, abundant semantic information in the multimedia data is mined, and vector distribution of different modal in the semantic space is fitted while ensuring that the types are unchanged and the modal can be distinguished. According to the method provided by the invention, the heterogeneity of different modal information can be effectively eliminated, and effective cross-media retrieval can be realized, thus, the method has wide market demand and application prospects in the fields of image and text retrieval, pattern recognition and the like.

(57) 摘要: 一种基于双语义空间的对抗性跨媒体检索方法, 涉及模式识别、自然语言处理、多媒体检索等技术领域; 包括: 特征生成过程、双语义空间的构建过程和对抗性语义空间优化过程。本发明通过建立同构双语义空间, 即文本子空间和图像子空间, 实现在最大限度保留原有图像和文本信息的同时消除语义鸿沟; 并通过对抗训练来优化同构子空间数据分布, 挖掘多媒体数据中丰富的语义信息, 在保证类别不变、模态可区分的情况下拟合语义空间中不同模态的向量分布。本发明方法能够有效的消除不同模态信息异构性, 实现有效的跨媒体检索, 在图文检索、模式识别等领域具有广泛的市场需求和应用前景。

一种基于双语义空间的对抗性跨媒体检索方法

技术领域

本发明涉及模式识别、自然语言处理、多媒体检索等技术领域，尤其涉及一种基于双语义空间的对抗性跨媒体检索方法，主要应用公共空间的特征映射来消除语义鸿沟，将不同模态的数据进行匹配达到检索的目的，并在跨媒体检索经典数据库中验证本方法的有效性。

背景技术

近年来，互联网技术飞速发展，随之而来的是多媒体信息的爆炸式增长，用户更倾向于通过检索获得多媒体信息结果，比如，用户输入关键字“狮子”，期望得到相关文字介绍以及其他模态的相关信息，比如狮子的图像，狮子的吼声以及狮子相关的视频等等。如此看来，传统的检索技术已经不能满足用户对于检索结果多样性、全面性的要求。跨媒体检索由于可以实现不同媒体间的灵活检索得到了广泛关注。其面临的挑战主要是不同模态的异构性和不可比性，以及不同模态所携带信息类型的不平衡性。比如，图片中包含更多的空间位置以及层次信息而文本中则包含着更多的上下文和背景信息。

现有方法大多将异构的特征映射到一个单一的同构空间，以消除“语义鸿沟”，但是，这样的处理同时伴随着大量的信息丢失，不同模态的特有信息不能得以保留，难以有效实现跨媒体检索。

发明内容

本发明提供了一种基于双语义空间的对抗性跨媒体检索方法，通过建立文本子空间和图像子空间，分别保留不同模态内部特征，并通过对抗训练来挖掘多媒体数据中丰富的语义信息，从而实现有效的跨媒体检索。

本发明的技术方案是：

一种基于双语义空间的对抗性跨媒体检索方法，通过建立文本子空间和图像子空间，分别保留不同模态内部特征，并通过对抗训练来挖掘多媒体数据中丰富的语义信息，从而实现有效的跨媒体检索；包括：特征生成过程、双语义空间的构建过程和对抗性语义空间优化过程。

1) 特征生成过程；

具体实施时，分别获取 NUS-WIDE-10k 和 Wikipedia 两个数据集的训练数据，验证数据及测试数据。并利用深度卷积神经网络 CNN (Convolutional Neural Network) 对训练及测试图像提取视觉特征向量，利用 BoW (Bag of Words) 模型对训练和测试文本提取“BoW 文本特

征向量”；设有 n 组训练数据，将图像和文本数据分别送入 CNN 网络和 BoW 模型，提取到的特征分别表示为图像特征 $I = \{i_1, i_2, \dots, i_n\}$ 和文本特征 $T = \{t_1, t_2, \dots, t_n\}$ ；

2) 双语义空间的构建过程，包括图像子空间和文本子空间；

本发明构建了双语义空间，即图像子空间和文本子空间。

21) 在图像子空间中，分别用三层的全连接网络实现对文本特征的映射以及图像特征在原有维度上的调整，同时引入三元组损失 (triplet loss) 进行同构空间特征优化，实现在最大限度保留图像信息的同时消除“语义鸿沟”。

具体执行如下操作：

211) 在图像子空间中，图像特征 I 经过三层全连接网络在原有维度 (4096 维) 上调整特征分布，为之后的图文匹配做准备，网络最后一层使用 ReLU 激活函数。同时引入一个相似的三层全连接网络将文本特征 T 映射到图像空间当中，尽量多的保留原有图像特征。

212) 在同构图像子空间中定义三元组 $\Phi^V = (v_i, t_j^+, t_k^-)$ ，其中 v_i 表示参照用图片， t_j^+ 为与 v_i 类别相同的正样例文本， t_k^- 表示与 v_i 类别不同的负样例文本，使用 L2 范数计算的不同模态数据间距离，表示为式 1：

$$l_2^V(v, t) = \|f_v(v) - f_t(t)\|_2^2 \quad (\text{式 1})$$

其中， $f_v(i)$ 与 $f_t(t)$ 分别为图像与文本的映射函数， $l_2^V(v, t)$ 表示同构空间内图像文本间的欧式距离。

以空间中的图像为参照，引入三元组约束 (triplet constraint) 调整文本分布，即拉近相同语义的文本图像对 (v_i, t_j^+) 的距离，同时增加语义不同的图像文本对 (v_i, t_k^-) 间距离。图像子空间三元组损失可以表示为式 2：

$$L_{tri}^V(\Phi^V) = \sum_{\Phi^V} \max(0, l_2^V(v_i, t_j^+) - l_2^V(v_i, t_k^-) + \alpha) \quad (\text{式 2})$$

其中， α 为表示安全系数的超参数； $L_{tri}^V(\Phi^V)$ 为图像子空间三元组损失； V 表示图像子空间。

22) 相似的，在文本子空间中实现图像特征映射以及文本特征调整。

具体执行如下操作：

221) 在文本子空间中，文本特征 T 利用文本映射函数 $\psi_t(T)$ 在原文本特征空间进行分布调整，图像映射函数 $\psi_v(I)$ 将图像特征 I 映射到文本空间当中，实现文本子空间内不同模态特征同构，同时避免大量损失原有文本信息。两个映射函数均为使用 ReLU 激活函数的 3 层全连

接网络。

222) 文本子空间的三元组损失 $L_{tri}^T(\Phi^T)$ 可以表示为式 3:

$$L_{tri}^T(\Phi^T) = \sum_{\Phi^T} \max(0, l_2^T(t_i, v_j^+) - l_2^T(t_i, v_k^-) + \alpha) \quad (\text{式 3})$$

其中, α 为表示安全系数的超参数, 与式 2 中相同。 Φ^T 表示文本子空间内的三元组 (v_i, t_j^+, t_k^-) , t_i 为参照用文本, v_j^+ 是与 t_i 类别相同的正样例图片, v_k^- 是与 t_i 类别不同的负样例图片。

223) 两个语义子空间平行工作, 实现特征提取的互补与平衡。

将双子空间的三元组损失适应性融合, 得到最终不同模态在同构空间新的特征分布, 表达式如式 4, 其中 η^V 、 η^T 为超参数:

$$L_{tri} = \eta^V \cdot L_{tri}^V(\Phi) + \eta^T \cdot L_{tri}^T(\Phi) \quad (\text{式 4})$$

3) 对抗性语义空间优化过程

本发明引入对抗性学习优化双语义空间, 在保证类别不变的前提下拟合不同模态的空间分布, 同时保证模态可判别, 上述结果反向传输更新双语义空间, 通过衡量空间距离完成“图像检索文本 (Img2Text)”和“文本检索图像 (Text2Img)”的跨媒体检索任务。

执行如下操作:

31) 采用类别预测 (concept prediction) 方法, 设定包含图像子空间和文本子空间的类别恒定目标函数, 分别对图像子空间和文本子空间进行优化, 使得子空间内不同模态特征映射或调整前后类别不变;

以图像子空间为例, 优化目标是最小化数据类别真实分布与同构空间内向量的类别预测分布间的误差, 其表达式如下:

$$L_c^V = -\frac{1}{N} \sum_{i=1}^N c_i \cdot (\log p_c(f_v(I)) + \log p_c(f_t(T))) \quad (\text{式 5})$$

式中, N 是一次迭代中数据的数目, c_i 为数据类别真实分布, $p_c(f_v(I))$ 、 $p_c(f_t(T))$ 分别为同构图像子空间内图像向量与文本向量的类别预测概率;

文本子空间内目标函数 L_c^T 定义类似:

$$L_c^T = -\frac{1}{N} \sum_{i=1}^N c_i \cdot (\log p_c(\psi_v(I)) + \log p_c(\psi_t(T))) \quad (\text{式 6})$$

$p_c(\psi_v(I))$ 与 $p_c(\psi_t(T))$ 分别为同构文本子空间内图像向量与文本向量的类别预测概率;

类别恒定目标函数整体可表示为式 7:

$$L_c = L_c^V + L_c^T \quad (\text{式 7})$$

32) 利用对抗性学习模型更新优化双空间网络参数, 其中将步骤 212)、221) 中的同构空间特征表示 $f_v(v)$ 、 $f_t(t)$ 、 $\psi_v(I)$ 与 $\psi_t(T)$ 作为对抗性学习模型的生成器生成向量, 生成器的优化目标为在保证类别恒定的前提下使得图像和文本在双子空间内的分布尽量相似, 优化函数表达式为式 8:

$$L_G = \mu_c \cdot L_c + \mu_{tri} \cdot L_{tri} \quad (\text{式 8})$$

其中, μ_c 、 μ_{tri} 为超参数。

对抗性学习模型的判别器的目标是尽量准确地判断出不同模态数据在双子空间内的向量是来自图像还是文本。具体实施时, 本发明使用 3 层前向传播全连接网络进行模态判别。

以图像子空间为例, 其优化目标是使得数据模态真实分布 m_i 与模态预测概率分布误差最小化, 其表达式为式 9:

$$L_D^V = -\frac{1}{N} \sum_{i=1}^N m_i \cdot (\log D(f_v(I)) + \log(1 - D(f_t(T)))) \quad (\text{式 9})$$

其中, N 是一次迭代中数据的数目, m_i 为数据真实模态, $D(f_v(I))$ 与 $D(f_t(T))$ 为同构图像子空间内图像与文本向量的模态预测。

文本子空间内目标函数 L_D^T 表达式与上述表达式类似:

$$L_D^T = -\frac{1}{N} \sum_{i=1}^N m_i \cdot (\log D(\psi_v(I)) + \log(1 - D(\psi_t(T)))) \quad (\text{式 10})$$

$D(\psi_v(I))$ 与 $D(\psi_t(T))$ 分别为同构文本子空间内图像与文本向量的模态预测。

判别器在双子空间内的目标函数整体可以表示为式 11:

$$L_D = L_D^V + L_D^T \quad (\text{式 11})$$

生成器与判别器的目标函数交替训练, 得到最优化的两个同构语义空间。

利用优化的同构双语义空间, 通过衡量空间距离完成“图像检索文本 (Img2Text)”和“文本检索图像 (Text2Img)”, 即可实现基于双语义空间的对抗性跨媒体的检索。

与现有技术相比, 本发明的有益效果是:

本发明提供了一种基于双语义空间的对抗性跨媒体检索方法, 其技术优势体现在:

(一) 通过建立文本子空间和图像子空间, 分别保留不同模态内部特征。在图像子空间中, 分别用三层的全连接网络实现对文本特征的映射以及图像特征在原有维度上的调整, 同

时引入三元组损失进行同构空间特征优化，实现在最大限度保留图像信息的同时消除“语义鸿沟”。相似的，在文本子空间中实现图像特征映射以及文本特征调整。两个语义子空间平行工作，实现特征提取的互补与平衡。

(二) 通过对抗训练来挖掘多媒体数据中丰富的语义信息，通过衡量空间距离完成“图像检索文本 (Img2Text)”和“文本检索图像 (Text2Img)”的跨媒体检索任务实现有效的跨媒体检索。

附图说明

图 1 是本发明提供方法的整体流程框图。

图 2 是本发明实施例中进行同构空间特征优化的示意图；

其中，(a) 是图像子空间内三元组分布优化；(b) 是文本子空间内三元组分布优化。

图 3 是本发明实施例进行文本检索图像的结果示例图，

其中，第一列为检索用文本，第二列为数据集给定的匹配图像，列 3 至列 7 为 mAP 值前五的对应检索结果。

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。

本发明提供了一种基于双语义空间的对抗性跨媒体检索方法，通过建立文本子空间和图像子空间，分别保留不同模态内部特征，并通过对抗训练来挖掘多媒体数据中丰富的语义信息，从而实现有效的跨媒体检索。

本发明提供的方法包括：特征生成过程、双语义空间的构建过程和对抗性语义空间优化过程；图 1 所示是本发明提供方法的流程，具体步骤如下：

- 1) 假设有 n 组训练数据，将图像和文本数据分别送入 CNN 网络和 BoW 模型，提取到的特征分别表示为图像特征 $I = \{i_1, i_2, \dots, i_n\}$ 和文本特征 $T = \{t_1, t_2, \dots, t_n\}$ ，
- 2) 在图像子空间中，图像特征 I 经过三层全连接网络在原有维度（4096 维）上调整特征分布，为之后的图文匹配做准备，网络最后一层使用 ReLU 激活函数。同时引入一个相似的三层全连接网络将文本特征 T 映射到图像空间当中，尽量多的保留原有图像特征。
- 3) 在同构图像子空间中定义三元组 $\Phi^V = (v_i, t_j^+, t_k^-)$ ，其中 v_i 表示参照用图片， t_j^+ 为与 v_i 类别相同的正样例文本， t_k^- 表示与 v_i 类别不同的负样例文本，使用 L2 范数计算的不同模

态数据间距离:

$$l_2^V(v, t) = \| f_v(v) - f_t(t) \|_2^2$$

其中 $f_v(i)$ 与 $f_t(t)$ 分别为图像与文本的映射函数, $l_2^V(v, t)$ 表示同构空间内图像文本间的欧式距离。

以空间中的图像为参照, 引入三元组约束 (triplet constraint) 调整文本分布, 即拉近相同语义的文本图像对 (v_i, t_j^+) 的距离, 同时增加语义不同的图像文本对 (v_i, t_k^-) 间距离。图像子空间三元组损失可以表示为:

$$L_{tri}^V(\Phi) = \sum_{\Phi} \max(0, l_2^V(v_i, t_j^+) - l_2^V(v_i, t_k^-) + \alpha)$$

其中 α 为表示安全系数的超参数。

- 4) 在文本子空间中, 文本特征 T 经过文本映射函数 $\psi_t(T)$ 在同原本文本维度空间特征分布调整, 图像映射函数 $\psi_v(I)$ 将图像特征 I 映射到文本空间当中, 实现文本子空间内不同模态特征同构, 同时避免大量损失原有文本信息。二者均为使用 ReLU 激活函数的 3 层全连接网络。与图像子空间类似, 文本子空间的三元组损失可以表示为:

$$L_{tri}^T(\Phi) = \sum_{\Phi} \max(0, l_2^T(t_i, v_j^+) - l_2^T(t_i, v_k^-) + \alpha)$$

其中 α 为表示安全系数的超参数。 Φ 表示文本子空间内的三元组 (v_i, t_j^+, t_k^-) , t_i 为参照用文本, v_j^+ 是与 t_i 类别相同的正样例图片, v_k^- 是与 t_i 类别不同的负样例图片。

- 5) 将双子空间的三元组损失适应性融合, 得到最终不同模态在同构空间新的特征分布, 表达式如下, 其中 η 为超参数:

$$L_{tri} = \eta^V \cdot L_{tri}^V(\Phi) + \eta^T \cdot L_{tri}^T(\Phi)$$

- 6) 引入类别预测 (concept prediction), 保证子空间内不同模态特征映射或调整前后类别不变, 以图像空间为例, 优化目标是最小化数据类别真实分布 c_i 与同构空间内向量的类别预测概率分布 $p_c(f_v(T))$ 、 $p_c(f_t(T))$ 间的误差, 其表达式如下:

$$L_c^V = -\frac{1}{N} \sum_{i=1}^N c_i \cdot (\log p_c(f_v(I)) + \log p_c(f_t(T)))$$

N 是一次迭代中数据的数目, c_i 为数据类别真实分布, $p_c(f_v(I))$ 、 $p_c(f_t(T))$ 分别为同

构图像子空间内图像向量与文本向量的类别预测概率。

文本子空间内目标函数 L_c^T 定义类似，类别恒定目标函数整体可表示为：

$$L_c = L_c^V + L_c^L$$

- 7) 对抗性学习模型更新优化双空间网络参数，其中步骤 3)、4) 中的同构空间特征表示 $f_v(v)$ 、 $f_t(t)$ 、 $\psi_v(I)$ 与 $\psi_t(T)$ 为对抗性学习模型中生成器生成向量，生成器的优化目标为在保证类别恒定的前提下使得图像和文本在双子空间内的分布尽量相似，优化函数表达式为：

$$L_G = \mu_c \cdot L_c + \mu_{tri} \cdot L_{tri}$$

其中 μ_c 、 μ_{tri} 为超参数。

判别器的目标是尽量准确地判断出不同模态数据在双子空间内的向量是来自图像还是文本。本发明使用 3 层前向传播全连接网络进行模态判别。以图像子空间为例，其优化目标是使得数据模态真实分布与模态预测概率分布误差最小化，其表达式为：

$$L_D^V = -\frac{1}{N} \sum_{i=1}^N m_i \cdot (\log D(f_v(I)) + \log(1 - D(f_t(T))))$$

其中， N 是一次迭代中数据的数目， m_i 为数据真实模态， $D(f_v(I))$ 与 $D(f_t(T))$ 为同构图像子空间内图像与文本向量的模态预测。

文本子空间内目标函数 L_D^T 表达式与上述表达式类似，判别器在双子空间内的目标函数整体可以表示为：

$$L_D = L_D^V + L_D^L$$

- 8) 生成器与判别器的目标函数交替训练，得到最优化的两个同构语义空间。利用优化的同构双语义空间，即可实现基于双语义空间的对抗性跨媒体的检索。

图 3 是本发明实施例中进行同构空间特征优化的示意图；其中，(a) 是图像子空间内三元组分布优化；(b) 是文本子空间内三元组分布优化。

表 1 给出了采用本发明提供的跨媒体检索方法及现有方法[1]-[6]在 NUS-WIDE-10k 和 Wikipedia 两个数据集进行检索得到的检索结果以得到的检索结果的对比。

表 1 NUS-WIDE-10k 和 Wikipedia 数据集在不同方法上的检索结果

方法	NUS-WIDE-10k 数据集			Wikipedia 数据集		
	图像检索 文本	文本检索 图像	平均	图像检索 文本	文本检索 图像	平均

CCA[1]	0.189	0.188	0.189	0.277	0.226	0.252
Multimodal DBN[2]	0.201	0.259	0.230	0.204	0.183	0.194
Corr-AE[3]	0.366	0.417	0.392	0.442	0.429	0.436
JRL[4]	0.426	0.376	0.401	0.479	0.428	0.454
ACMR[5]	0.544	0.538	0.541	0.509	0.431	0.470
MCSM[6]	—	—	—	0.516	0.458	0.487
本发明	0.569	0.565	0.567	0.538	0.475	0.507

其中，现有方法[1]–[6]分别为：

文献[1]（Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: International Conference on Multimedia. pp. 251–260 (2010)）记载的 CCA 方法；

文献[2]（Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: ICML Workshop）记载的 Multimodal DBN 方法；

文献[3]（Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder pp. 7–16 (2014)）记载的 Corr-AE 方法；

文献[4]（Zhai, X., Peng, Y., Xiao, J.: Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Transactions on Circuits and Systems for Video Technology 24(6), 965–978 (2014)）记载的 JRL 方法；

文献[5]（Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: ACM on Multimedia Conference. pp. 154–162 (2017)）记载的 ACMR 方法；

文献[6]（Peng, Y., Qi, J., Yuan, Y.: Modality-specific cross-modal similarity measurement with recurrent attention network (2017)）记载的 MCSM 方法。

表 2 给出了本发明提供的跨媒体检索方法及本发明两个变体（仅有图像空间和仅有文本空间）对 NUS-WIDE-10k 和 Wikipedia 两个数据集进行检索，得到的检索结果的对比。

表 2 NUS-WIDE-10k 和 Wikipedia 数据集在图像空间、文本空间和双语义空间的检索结果

方法	NUS-WIDE-10k 数据集			Wikipedia 数据集		
	图像检索 文本	文本检索 图像	平均	图像检索 文本	文本检索 图像	平均
只有图像空间	0.572	0.516	0.544	0.527	0.465	0.496
只有文本空间	0.582	0.494	0.538	0.531	0.465	0.498
双语义空间	0.569	0.565	0.567	0.538	0.475	0.507

表 1 和表 2 中，检索结果用 mAP 值衡量，mAP 值越高，检索效果越优异。

从表 1 中可以看出，与现有方法比较，本发明在图像检索文本和文本检索图像两大任务上检索正确率均有明显提升，表 2 结果显示，虽然 NUS-WIDE-10k 数据集上的检索结果显示，在图像检索文本任务上双语义空间的 mAP 值稍低于仅在文本空间的结果，但双空间的平均检索结果在两个数据集上均明显高于单空间，充分验证了本发明中所提出的双语义空间检索结构的有效性。图 3 是本发明实施例进行文本检索图像的结果示例图。图中，第一列为检索用文本，第二列为数据集给定的匹配图像，列 3 至列 7 为 mAP 值前五的对应检索结果，显示了本发明用于检索的有效性。

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、一种基于双语义空间的对抗性跨媒体检索方法，通过建立文本子空间和图像子空间，分别保留不同模态内部特征，并通过对抗训练挖掘多媒体数据中的语义信息，从而实现有效的跨媒体检索；包括：特征生成过程、双语义空间的构建过程和对抗性语义空间优化过程；

1) 特征生成过程；执行如下操作：

获取训练数据，验证数据及测试数据；

利用深度卷积神经网络 CNN 模型，对训练及测试图像提取视觉特征向量，利用 BoW 模型对训练和测试文本提取得到 BoW 文本特征向量；设有 n 组训练数据，提取到的特征分别表示为图像特征 $I = \{i_1, i_2, \dots, i_n\}$ 和文本特征 $T = \{t_1, t_2, \dots, t_n\}$ ；

2) 双语义空间的构建过程，包括图像子空间和文本子空间；执行如下操作：

21) 在图像子空间中，分别用三层的全连接网络实现对文本特征的映射以及图像特征在原有维度上的调整，同时引入三元组损失进行同构空间特征优化，保留图像信息并消除语义鸿沟；具体执行如下操作：

211) 在图像子空间中，图像特征 I 经过三层全连接网络在原有维度上调整特征分布，网络最后一层使用 ReLU 激活函数；引入三层全连接网络，将文本特征 T 映射到图像空间当中，保留原有图像特征；

212) 在同构图像子空间中定义三元组 $\Phi^V = (v_i, t_j^+, t_k^-)$ ，其中 v_i 表示参照用图片， t_j^+ 为与 v_i 类别相同的正样例文本， t_k^- 表示与 v_i 类别不同的负样例文本；使用 L2 范数计算的不同模态数据间距离，表示为式 1：

$$l_2^V(v, t) = \|f_v(v) - f_t(t)\|_2^2 \quad (\text{式 1})$$

其中， $f_v(i)$ 与 $f_t(t)$ 分别为图像与文本的映射函数， $l_2^V(v, t)$ 表示同构空间内图像文本间的欧式距离；

引入图像子空间三元组约束调整文本图像分布，拉近相同语义的文本图像对 (v_i, t_j^+) 的距离，同时增加语义不同的图像文本对 (v_i, t_k^-) 间距离；图像子空间三元组损失 $L_{tri}^V(\Phi^V)$ 可以表示为式 2：

$$L_{tri}^V(\Phi^V) = \sum_{\Phi^V} \max(0, l_2^V(v_i, t_j^+) - l_2^V(v_i, t_k^-) + \alpha) \quad (\text{式 2})$$

其中， α 为表示安全系数的超参数； $L_{tri}^V(\Phi^V)$ 为图像子空间三元组损失； V 表示图像子空间；

22) 在文本子空间中实现图像特征映射及文本特征调整；执行如下操作：

221) 在文本子空间中，文本特征 T 利用文本映射函数 $\psi_t(T)$ 在原文本特征空间进行分布调整，图像映射函数 $\psi_v(I)$ 将图像特征 I 映射到文本空间当中，实现文本子空间内不同模态特征同构；

222) 文本子空间的三元组损失 $L_{tri}^T(\Phi^T)$ 表示为式 3：

$$L_{tri}^T(\Phi^T) = \sum_{\Phi^T} \max(0, l_2^T(t_i, v_j^+) - l_2^T(t_i, v_k^-) + \alpha) \quad (\text{式 3})$$

其中， α 为表示安全系数的超参数； Φ^T 表示文本子空间内的三元组 (v_i, t_j^+, t_k^-) ， t_i 为参照用文本， v_j^+ 是与 t_i 类别相同的正样例图片， v_k^- 是与 t_i 类别不同的负样例图片；

224) 图像子空间和文本子空间平行工作，实现特征提取的互补与平衡；

将双子空间的三元组损失适应性融合，得到最终不同模态在同构空间新的特征分布，表示为式 4，其中 η^V 、 η^T 为超参数：

$$L_{tri} = \eta^V \cdot L_{tri}^V(\Phi) + \eta^T \cdot L_{tri}^T(\Phi) \quad (\text{式 4})$$

3) 对抗性语义空间优化过程；执行如下操作：

31) 采用类别预测方法，设定包含图像子空间和文本子空间的类别恒定目标函数，分别对图像子空间和文本子空间进行优化，使得子空间内不同模态特征映射或调整前后类别不变；

设定图像子空间内的目标函数 L_c^V 、文本子空间内的目标函数 L_c^T 是最小化数据类别真实分布与同构空间内向量的类别预测分布间的误差；类别恒定目标函数表示为式 7：

$$L_c = L_c^V + L_c^T \quad (\text{式 7})$$

32) 利用对抗性学习模型更新优化双空间网络参数，将步骤 212)、221) 中的同构空间特征表示 $f_v(v)$ 、 $f_t(t)$ 、 $\psi_v(I)$ 与 $\psi_t(T)$ 作为对抗性学习模型的生成器生成向量，生成器的优化目标为在保证类别恒定的前提下使得图像和文本在双子空间内的分布尽量相似，优化函数 L_G 表达式为式 8：

$$L_G = \mu_c \cdot L_c + \mu_{tri} \cdot L_{tri} \quad (\text{式 8})$$

其中， μ_c 、 μ_{tri} 为超参数；

对抗性学习模型的判别器的目标是尽量准确地判断出不同模态数据在双子空间内的向量是来自图像还是文本；

设定图像子空间内的优化目标为 L_D^V ，文本子空间内目标函数为 L_D^T ，判别器在双子空间内

的目标函数 L_D 表示为式 11:

$$L_D = L_D^V + L_D^T \quad (\text{式 11})$$

生成器与判别器的目标函数交替训练, 得到最优化的两个同构语义空间;

利用最优化的同构双语义空间, 完成“图像检索文本 (Img2Text)”和“文本检索图像 (Text2Img)”, 即可实现基于双语义空间的对抗性跨媒体的检索。

2、如权利要求 1 所述基于双语义空间的对抗性跨媒体检索方法, 其特征是, 步骤 31) 中, 对图像子空间, 优化目标是最小化数据类别真实分布与同构空间内向量的类别预测分布间的误差, 表示为式 5:

$$L_c^V = -\frac{1}{N} \sum_{i=1}^N c_i \cdot (\log p_c(f_v(I)) + \log p_c(f_t(T))) \quad (\text{式 5})$$

式中, N 是一次迭代中数据的数目, c_i 为数据类别真实分布, $p_c(f_v(I))$ 、 $p_c(f_t(T))$ 分别为同构图像子空间内图像向量与文本向量的类别预测概率;

文本子空间内目标函数 L_c^T 定义为式 6:

$$L_c^T = -\frac{1}{N} \sum_{i=1}^N c_i \cdot (\log p_c(\psi_v(I)) + \log p_c(\psi_t(T))) \quad (\text{式 6})$$

其中, $p_c(\psi_v(I))$ 与 $p_c(\psi_t(T))$ 分别为同构文本子空间内图像向量与文本向量的类别预测概率。

3、如权利要求 1 所述基于双语义空间的对抗性跨媒体检索方法, 其特征是, 步骤 32) 中, 对图像子空间, 优化目标是使得数据模态真实分布 m_i 与模态预测概率分布误差最小化, 表示为式 9:

$$L_D^V = -\frac{1}{N} \sum_{i=1}^N m_i \cdot (\log D(f_v(I)) + \log (1 - D(f_t(T)))) \quad (\text{式 9})$$

其中, N 是一次迭代中数据的数目, m_i 为数据真实模态, $D(f_v(I))$ 与 $D(f_t(T))$ 为同构图像子空间内图像与文本向量的模态预测;

文本子空间内目标函数 L_D^T 表示为式 10:

$$L_D^T = -\frac{1}{N} \sum_{i=1}^N m_i \cdot (\log D(\psi_v(I)) + \log (1 - D(\psi_t(T)))) \quad (\text{式 10})$$

$D(\psi_v(I))$ 与 $D(\psi_t(T))$ 分别为同构文本子空间内图像与文本向量的模态预测。

4、如权利要求 1 所述基于双语义空间的对抗性跨媒体检索方法, 其特征是, 步骤 32) 具体使用 3 层前向传播全连接网络进行模态判别。

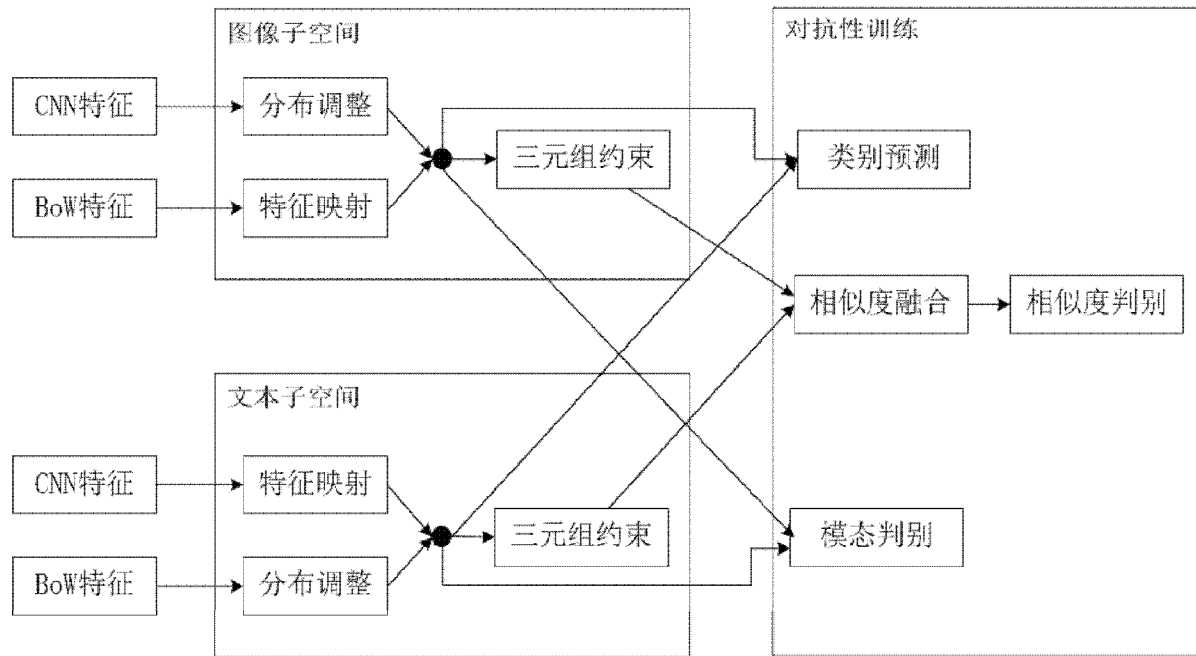
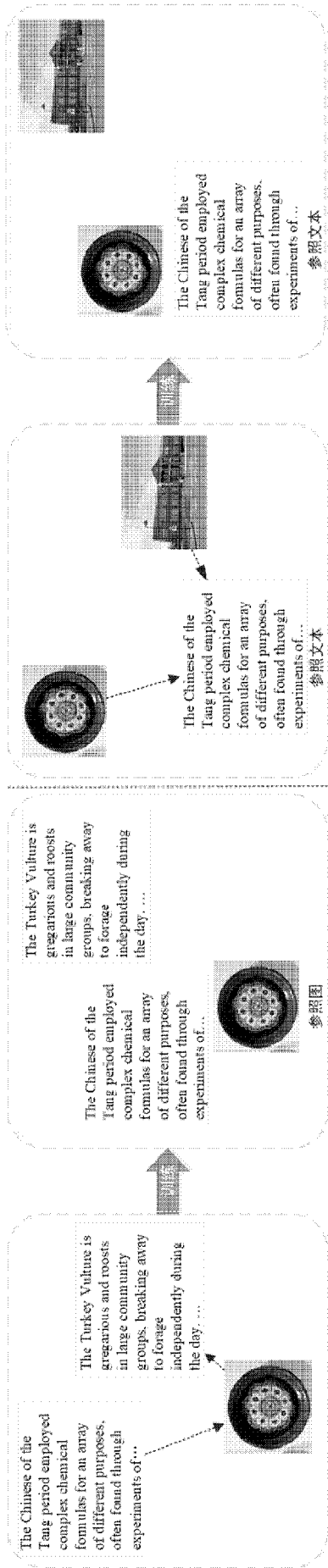


图 1

16. 5月 2019(16. 05. 2019)



(b) 文本子空间内三元组分布优化

(a) 图像子空间内三元组分布优化

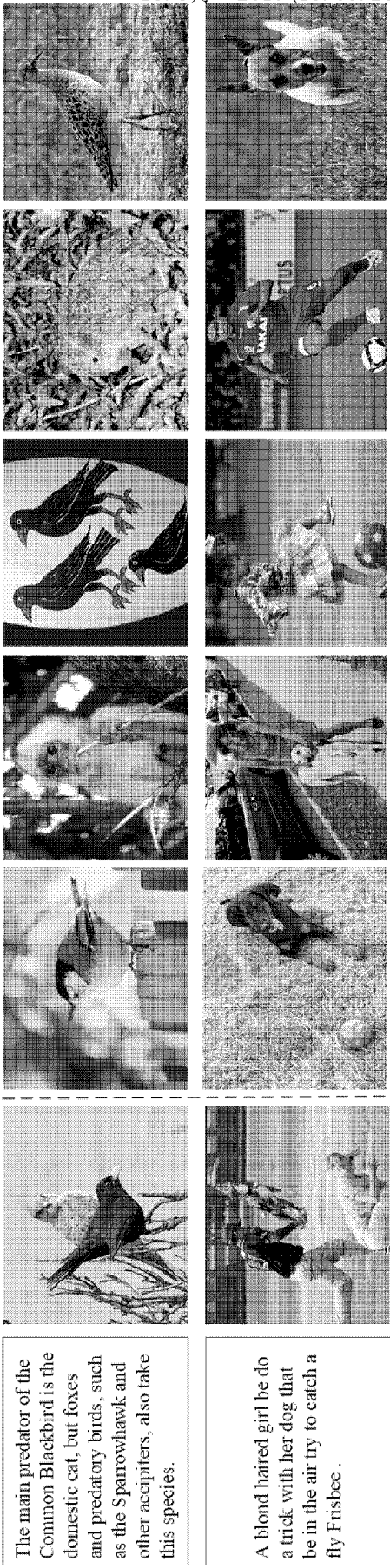


图 2

图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/076400

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/48(2019.01)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

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)

WPI, EPODOC, CNPAT, CNKI: 文本, 图像, 媒体, 检索, TEXT, IMAGE, MEDIUM, SEARCH

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109344266 A (PEKING UNIVERSITY SHENZHEN GRADUATE SCHOOL) 15 February 2019 (2019-02-15) entire document	1-4
A	CN 107562812 A (PEKING UNIVERSITY) 09 January 2018 (2018-01-09) entire document	1-4
A	CN 104317834 A (ZHEJIANG UNIVERSITY) 28 January 2015 (2015-01-28) entire document	1-4
A	CN 106095829 A (HUAQIAO UNIVERSITY) 09 November 2016 (2016-11-09) entire document	1-4



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

21 March 2019

Date of mailing of the international search report

03 April 2019

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/076400

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109344266	A	15 February 2019	None			
CN	107562812	A	09 January 2018	None			
CN	104317834	A	28 January 2015	CN	104317834	B	29 September 2017
CN	106095829	A	09 November 2016	None			

国际检索报告

国际申请号

PCT/CN2019/076400

A. 主题的分类

G06F 16/48(2019.01)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

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))

WPI, EPDOC, CNPAT, CNKI: 文本, 图像, 媒体, 检索, TEXT, IMAGE, MEDIUM, SEARCH

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 109344266 A (北京大学深圳研究生院) 2019年 2月 15日 (2019 - 02 - 15) 全文	1-4
A	CN 107562812 A (北京大学) 2018年 1月 9日 (2018 - 01 - 09) 全文	1-4
A	CN 104317834 A (浙江大学) 2015年 1月 28日 (2015 - 01 - 28) 全文	1-4
A	CN 106095829 A (华侨大学) 2016年 11月 9日 (2016 - 11 - 09) 全文	1-4

☐ 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

2019年 3月 21日

国际检索报告邮寄日期

2019年 4月 3日

ISA/CN的名称和邮寄地址

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

传真号 (86-10)62019451

受权官员

林亮亮

电话号码 86-(010)-62411890

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/076400

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	109344266	A	2019年 2月 15日	无	
CN	107562812	A	2018年 1月 9日	无	
CN	104317834	A	2015年 1月 28日	CN 104317834	B 2017年 9月 29日
CN	106095829	A	2016年 11月 9日	无	