arxiv 1702.05729

`trans`

Facing the limitations of both modalities, we propose to use natural language description to search person. It does not require a person photo to be given as in those image- based query methods. Natural language also can precisely describe the details of person appearance, and does not re- quire labelers to go through the whole list of attributes.

面对这两种模式的局限性，我们建议使用自然语言描述搜索人。它不需要像在基于图像的查询方法中那样提供person照片。自然语言也可以精确地描述人的外貌细节，而不需要重新标记整个属性列表。

Since there is no existing dataset focusing on describ- ing person appearances with natural language, we first build a large-scale language dataset, with 40,206 images of 13,003 persons from existing person re-identification datasets. Each person image is described with two sen- tences by two independent workers on Amazon Mechan- ical Turk (AMT). On the visual side, the person images pooled from various re-identification datasets are under dif- ferent scenes, view points and camera specifications, which increases the image content diversity. On the language side, the dataset has 80,412 sentence descriptions, contain- ing abundant vocabularies, phrases, and sentence patterns and structures. The labelers have no limitations on the lan- guages for describing the persons. We perform a series of user studies on the dataset to show the rich expression of the language description. Examples from the dataset are shown in Figure 2.

由于没有现有的数据集聚焦于描述人的自然语言现象，我们首先构建了一个大规模的语言数据集，从现有的人员重新识别数据集中，有40206张图片，13003人。亚马逊Mechan- ical Turk (AMT)上的两位独立工作者用两种不同的方式描述每个人的形象。在视觉方面，从不同的再识别数据集中汇集出来的人物图像在不同的场景、视图点和相机规格下，增加了图像内容的多样性。在语言方面，数据集有80412个句子描述，包含丰富的词汇、短语、句型和结构。标签对描述人的局域网语言没有限制。我们对数据集进行了一系列的用户研究，以显示语言描

述的丰富表达。数据集中的示例如图2所示。

We propose a novel Recurrent Neural Network with Gated Neural Attention (GNA-RNN) for person search. The GNA-RNN takes a description sentence and a person image as input and outputs the affinity between them. The sentence is input into a word-LSTM and processed word by word. At each word, the LSTM generates unit-level attentions for individual visual units, each of which deter- mines whether certain person semantic attributes or visual patterns exist in the input image. The visual-unit attention mechanism weights the contributions of different units for different words. In addition, we also learn word-level gates that estimate the importance of different words for adap-tive word-level weighting. The final affinity is obtained by averaging over all units' responses at all words. Both the unit-level attention and word-level sigmoid gates contribute to the good performance of our proposed GNA-RNN.

我们提出了一种新的具有门控神经注意(GNA-RNN)的再发神经网络用于人的搜索。GNA-RNN以描述句和人物形象作为输入并输出它们之间的关联性。句子被输入一个单词，然后逐字处理。在每个单词中，LSTM会对单个的视觉单元产生单元级的关注，每个视觉单元都能阻止——在输入图像中是否存在某些人的语义属性或视觉模式。视觉单位注意机制衡量不同单位对不同词汇的贡献。此外，我们还学习了单词级的门，它估计了不同单词对适应词级权重的重要性。最终的亲和力是通过对所有单位在所有词语上的反应进行平均得到的。单元级的注意和字级的sigmoid gates都有助于我们建议的GNA-RNN的良好性能。

The contribution of this paper is three-fold. 1) We pro- pose to study the problem of searching persons with nat- ural language. This problem setting is more practical for real-world scenarios. To support this research direction, a large-scale person description dataset with rich language annotations is collected and the user study on the natural language description of person is given. 2) We investigate a wide range of plausible solutions based on different vi- sion and language frameworks, including image caption- ing [19, 37], visual QA [45, 32], and visual-semantic em- bedding [31], and establish baselines on the person search benchmark. 3) We further propose a novel Recurrent Neu- ral Network with Gated Neural Attention (GNA-RNN) for person search, with the state-of-the-art performance on the person search benchmark.

本文的贡献有三方面。1)我们主张研究用自然语言搜索人的问题。这个问题设置在实际场景中

更实用。为了支持这一研究方向，收集了具有丰富语言标注的大型人物描述数据集，并对人物的自然语言描述进行了用户研究。2)基于不同的vi- sion和语言框架，我们研究了多种可行的解决方案，包括图片标题- ing [19, 37]， visual QA[45, 32]，可视化-语义em-层理[31]，并在person搜索基准上建立基线。3)我们进一步提出了一种新型的神经网络(GNA-RNN)，用于人的搜索，在人的搜索基准上具有最先进的性能。

1.1. Related work

As there are no existing datasets and methods designed for the person search with natural language, we briefly sur- vey the language datasets for various vision tasks, along with the deep language models for vision that can be used as possible solutions for this problem.

由于没有为使用自然语言进行搜索而设计的现有数据集和方法，因此我们简要地确定了用于各种视觉任务的语言数据集，以及可作为解决此问题的可能解决方案的深度语言模型。

Language datasets for vision. Early language datasets for vision include Flickr8K [12] and Flickr30K [42]. In- spired by them, Chen et al. built a larger MS-COCO Cap- tion [2] dataset. They selected 164,062 images from MS- COCO [25] and labeled each image with five sentences from independent labelers. Recently, Visual Genome [20] dataset was proposed by Krishna et al., which incorporates dense annotations of objects, attributes, and relationships within each image. However, although there are persons in the datasets, they are not the main subjects for descriptions and cannot be used to train person search algorithms with language descriptions. For fine-grained visual descriptions, Reed et al. added language annotations to Caltech-UCSD birds [38] and Oxford-102 flowers [29] datasets to describe contents of images for text-image joint embedding.

对视觉语言的数据集。早期的视觉语言数据集包括Flickr8K[12]和Flickr30K[42]。在他们的启发下，陈等人建立了一个更大的MS-COCO Cap-tion[2]数据集。他们从MS-COCO[25]中选取了164,062张图片，并在每张图片上贴上独立标签者的五句话。最近，Krishna等人提出了Visual Genome[20]数据集，该数据集包含了每个图像中对象、属性和关系的密集注释。然而，尽管数据集中有个人，但他们不是描述的主要主体，不能用语言描述训练人搜索算法。对于细粒度的视觉描述，Reed等人在Caltech-UCSD birds[38]和Oxford-102flowers[29]数据

集中添加了语言注释，用于描述图像内容，用于文本图像拼接嵌入。

Deep language models for vision. Different from con- volutional neural network which works well in image clas- sification [21, 10] and object detection [18, 17, 16], re- current neural network is more suitable in processing se- quential data. A large number of deep models for vision tasks [40, 1, 13, 15, 8, 3, 5] have been proposed in recent years. For image captioning, Mao et al. [28] learned feature embedding for each word in a sentence, and connected it with the image CNN features by a multi- modal layer to gen- erate image captions. Vinyal et al. [37] extracted high-level image features from CNN and fed it into LSTM for estimat- ing the output sequence. The NeuralTalk [19] looked for the latent alignment between segments of sentences and image regions in a joint embedding space for sentence generation.

深度语言模型。与适用于图像clas-sification[21,10]和对象检测[18,17,16]的convolutional neural network不同，recurrent neural network更适合处理se-quential次元数据。近年来，人们提出了大量的视觉任务深度模型[40,1,13,15,8,3,5]。对于图像字幕，Mao等人在一个句子中学习了每个单词的特征嵌入，并通过多模态层将其与CNN的图像特征连接到生成图像字幕。Vinyal et al.[37]从CNN提取高级图像特征，输入LSTM估计输出序列。神经语言[19]在句子生成的联合嵌入空间中寻找句子片段和图像区域之间的潜在对齐。

Visual QA methods were proposed to answer questions about given images [32, 30, 41, 34, 27, 7]. Yang et al. [41] presented a stacked attention network that refined the joint features by recursively attending question-related image regions, which leads to better QA accuracy. Noh et al. [30]learned a dynamic parameter layer with hashing techniques,which adaptively adjusts image features based on different questions for accurate answer classification.

提出了Visual QA方法来回答给定图像的问题[32、30、41、34、27、7]。Yang等人提出了一种叠加式的注意网络，通过递归地处理与问题相关的图像区域来细化联合特征，从而提高了QA的准确性。Noh等人利用哈希技术学习了一个动态参数层，基于不同的问题自适应地调整图像特征，从而实现准确的回答分类。

Visual-semantic embedding methods [6, 19, 31, 26, 33] learned to embed both language and images into a common space for image classification and retrieval. Reed et al. [31] trained an end-to-end CNN-RNN model which jointly embeds the

images and fine-grained visual descriptions into the same feature space for zero-shot learning. Text-to-image retrieval can be conducted by calculating the distances in the embedding space. Frome et al. [6] associated seman- tic knowledge of text with visual objects by constructing a deep visual-semantic model that re-trained the neural lan- guage model and visual object recognition model jointly.

视觉-语义嵌入方法[6,19,31,26,33]学会将语言和图像嵌入到一个公共空间中，用于图像分类和检索。Reed等人训练了一个端到端cnn- rnn模型，该模型将图像和细粒度的视觉描述共同嵌入到相同的特征空间中，以便进行零距离学习。文本到图像的检索可以通过计算嵌入空间的距离来实现。Frome et al.[6]通过构建深度视觉-语义模型，将文本的语义知识与视觉对象相关联，对神经language模型和视觉对象识别模型进行联合训练。

2. Benchmark for person search with natural language description

Since there is no existing language dataset focusing on person appearance, we build a large-scale benchmark for person search with natural language, termed as CUHK Person Description Dataset (CUHK-PEDES). We collected 40,206 images of 13,003 persons from five existing person re-identification datasets, CUHK03 [23], Market-1501 [43], SSM [39], VIPER [9], and CUHK01 [22], as the subjects for language descriptions. Since persons in Market-1501 and CUHK03 have many similar samples, to balance the number of persons from different domains, we randomly selected four images for each person in the two datasets. All the image were labeled by crowd workers from Amazon Mechanical Turk (AMT), where each image was annotated with two sentence descriptions and a total of 80,412 sen- tences were collected. The dataset incorporates rich details about person appearances, actions, poses and interactions with other objects. The sentence descriptions are generally long (> 23 words in average), and has abundant vocabulary and little repetitive information. Examples of our proposed dataset are shown in Figure 2.

由于目前还没有以人的外貌为中心的语言数据集，所以我们为使用自然语言的人的搜索建立了一个大规模的基准，称为中大人的描述数据集(CUHK- pedes)。我们收集了来自5个现有人员再识别数据集、CUHK03[23]、Market-1501[43]、SSM[39]、VIPER[9]和CUHK01[22]的

40,206张13003人的图像作为语言描述的对象。由于Market-1501和CUHK03中的人有很多相似的样本，为了平衡来自不同领域的人的数量，我们在两个数据集中为每个人随机选择了4张图像。所有的图像都被来自亚马逊土耳其机器人(AMT)的人群贴上了标签，每一张图片都有两个句子描述，总共收集了80,412个句子。该数据集包含关于人的外观、动作、姿势和与其他对象的交互的丰富细节。句子描述一般较长(平均为>23个单词)，词汇量丰富，重复信息较少。我们建议的数据集的示例如图2所示。

2.1. Dataset statistics

The dataset consists of rich and accurate annotations with open word descriptions. There were 1,993 unique workers involved in the labeling task, and all of them have greater-than 95% approving rates. We asked the workers to describe all important characteristics in the given images using sentences with at least 15 words. The large number of workers means the dataset has diverse language descrip- tions and methods trained with it are unlikely to overfit to descriptions of just a few workers.

数据集由丰富而准确的注释和开放的词描述组成。有1993个工人参与了标签任务，他们都有超过95%的批准率。我们要求工作人员用至少15个词的句子描述给定图像中的所有重要特征。大量的工作人员意味着数据集具有不同的语言描述和训练过的方法，不太可能超出对少数工作人员的描述。

Vocabulary, phrase sizes, and sentence length are impor- tant indicators on the capacity our language dataset. There are a total of 1,893,118 words and 9,408 unique words in our dataset. The longest sentence has 96 words and the av- erage word length is 23.5 which is significantly longer than the 5.18 words of MS-COCO Caption [25] and the 10.45 words of Visual Genome [20]. Most sentences have 20 to 40 words in length. Figure 3 illustrates some person exam- ples and high-frequency words.

词汇、短语大小和句子长度是影响语言数据集容量的重要指标。在我们的数据集中总共有1,893,118个单词和9,408个唯一的单词。最长的句子有96个单词，av- erage单词长度为23.5个，明显比MS-COCO标题[25]的5.18个单词和视觉基因组[20]的10.45个单词长。大多数句子有20到40个单词。图3展示了一些人的测试词和高频词。

2.2. User study

Based on the language annotations we collect, we con- duct the user studies to investigate 1) the expressive power of language descriptions compared with that of attributes, 2) the expressive power in terms of the number of sentences and sentence length, and 3) the expressive power of different word types. The studies provide us insights for understanding the new problem and guidance on designing our neural networks.

基于我们收集的语言注释，我们对用户进行了研究:1)语言描述相对于属性的表达能力;2)句子数量和句子长度的表达能力;3)不同类型词的表达能力。这些研究为我们理解新问题和设计神经网络提供了指导。

Language vs. attributes. Given a descriptive sentence or annotated attributes of a query person image, we ask crowd workers from AMT to select its corresponding image from a pool of 20 images. The 20 images consist of the ground truth image, 9 images with similar appear- ances to the ground truth, and 10 randomly selected images from the whole dataset. The 9 similar images are chosen from the whole dataset by the LOMO+XQDA [24] method, which is a state-of-the-art method for person reidentification. The other 10 distractor images are randomly selected and have no overlap with the 9 similar images. The person attribute annotations are obtained from the PETA [4] dataset, which have 1,264 same images with our dataset. A total of 500 images are manually searched by the workers, and the average top-1 and top-5 accuracies of the searches are evaluated. The searches with language descriptions have 58.7% top-1 and 92.0% top-5 accuracies, while the searches with attributes have top-1 and top-5 accuracies of 33.3% and 74.7% respectively. In terms of the average used time for each search, using language descriptions takes 62.18s, while using attributes takes 81.84s. The results show that, from human's perspective, language descriptions are much precise and effective in describing persons than attributes. They partially endorse our choice of using language de- scriptions for person search.

语言与属性。给定一个描述性句子或查询人映像的带注释属性，我们要求AMT的工作人员从20个映像池中选择相应的映像。这20幅图像由地面真相图像、9幅与地面真相相似的图像以及整个数据集随机选取的10幅图像组成。通过LOMO+XQDA[24]方法从整个数据集中选择9幅相似的图像，这是一种最先进的人员再识别方法。其他10个干扰图像是随机选择的，与9个相

似图像没有重叠。person属性注释来自PETA[4]数据集，该数据集有1264个与我们的数据集相同的图像。工作人员手工搜索了总共500幅图像，并对搜索结果的前5位和前5位的平均精度进行了评估。使用语言描述的搜索有58.7%的top-1和92.0%的top-5准确性，而使用属性的搜索有33.3%和74.7%的top-1和top-5准确性。就每次搜索的平均使用时间而言，使用语言描述需要62.18秒，而使用属性需要81.84秒。结果表明，从人的角度来看，语言描述比属性更准确、更有效地描述人。他们在一定程度上赞成我们选择用文字搜索人。

Sentence number and length. We design manual ex- periments to investigate the expressive power of language descriptions in terms of the number of sentences for each image and sentence length. The images in our dataset are categorized into different groups based on the number of sentences associated with each image and based on different sentence lengths. Given the sentences for each image, we ask crowd workers from AMT to manually retrieve the corresponding images from pools of 20 images. The average top-1 and top-5 accuracies, and used time for different image groups are shown in Figure 4, which show that 3 sentences for describing a person achieved the highest retrieval accuracy. The longer the sentences are, the easier for users to retrieve the correct images.

句子的数量和长度。我们设计了一种手工辅助工具来研究语言描述在每个图像的句子数量和句子长度方面的表达能力。我们数据集中的图像根据与每个图像相关的句子数量和不同的句子长度被分为不同的组。给定每个图像的句子，我们要求AMT的工作人员从20个图像池中手动检索相应的图像。年龄最大的1和前5个准确率以及对不同图像组的使用时间如图4所示，其中3个描述一个人的句子的检索准确率最高。句子越长，用户越容易检索到正确的图像。