指导教师：杨涛　　提交时间：2016/3/17

CVPR2015 Paper Translation

No: 01

姓名：尹宁

学号：2013302469

班号：10011301

多段查询指令中图像序列的排序和检索

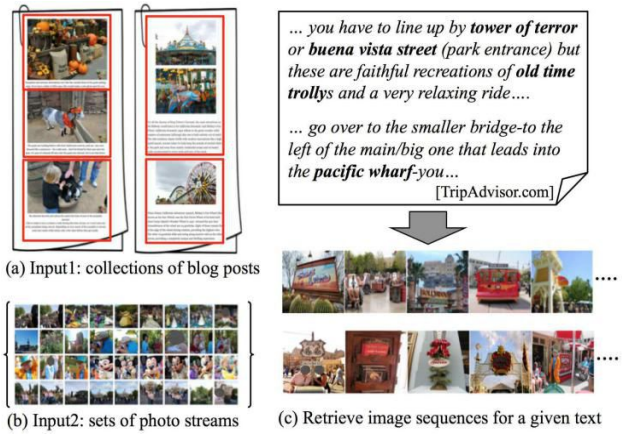
Gunhee Kim
Seoul National University
gunhee@cs.cmu.edu

Seungwhan Moon
Carnegie Mellon University
seungwhm@cs.cmu.edu

Leonid Sigal
Disney Research Pittsburgh
lsigal@disneyresearch.com

摘要

我们提出一种方法：从一个自然语言的文本（包括多个句子或段落）查询中进行图像序列的排序和检索。这种方法的关键应用之一是通过自动检索最具说明性的描叙图片的句子，使参观者在 Tripadvisor（全球旅游网站）和 Yelp（美国的点评网站）的文字评论可视化。最之前的工作是解决自然语言的句子和图片或者视频的关系，我们的工作还扩展到了文字段落和图片序列的关系。我们的这种方法是利用用户在网上发布博客文章和照片流产生的庞大的资源。我们利用共同部署的信息文本的博客文章和用户精心挑选出来的具有代表性的图片作为图文并行训练的数据。我们开发了大规模的图片流来增加图片检索的样本。我们还设计了一个潜在结构性的支持向量机架构来学习文本和图像序列的语义相关关系。我们提出了对于新创建的迪斯尼数据集的定量和定性的结果。



1.概述

文本的和可视化的这两种形式在交流中的许多方面是互补和协作的关系。（比如：新闻文章和博客）。能够利用自由形式的文本并能自动给它们配上相关图片的系统，将会成为往联合理解自然语言描述和图像视觉内容方向发展的重要一步。我们最近一直在稳步的解决这个具有挑战性的问题：对图像描述的一句话（即主-谓-宾风格）的文本产生方法【7, 8, 9, 17, 24, 29】或者在句子查询指令中的图像、视频检索【10, 37】。在本论文中，我们下一个飞跃将是用图像序列的检索（而不是单个图像）来阐述更长的内容，就是那种可能包含多个句子甚至段落（不是单个句子）的文本文章。挑战之一是，图 1 所示：一个描述迪斯尼乐园的问题示例。我们利用（a）中多的博客文章来学习句子和图像之间的映射关系。（b）中的照片流来增加图像样本。（c）给定的包括多句或多段落的文本查询指令会产生最确切描述的图片序列，我们的目标

是对这些图片序列进行排序和检索。

然而，从文本和图像的语义关系中获得适当的文本-图像平行语料库是可以实现的。

随着社会媒体站点的激增，个人自愿地在大量的网络平台上以图片、视频或文字的方式来分享他们自己的经历。例如，许多人在参观过迪斯尼乐园后拍下了大量关于他们独特经历的照片流并把这些照片上传到照片保存站点，比如 Flickr。还有一些更热情的用户会在微博上发表文章来记录旅行，比如 Blogger 和 WordPress，或者在评论网站上评价他们的旅行，比如 Tripadvisor 和 Yelp。

在本论文中，我们将通过利用大型语料库的在线图像和以不同形式描述常见事件和活动的文本，来解决对多句自然语言查询指令的图像序列的排序和检索这一问题。图 1 用一个旅游的例子阐述了我们的问题陈述的直觉（例如，参观迪斯尼乐园）。给定一个由多个句子或段落组成的文本查询指令，我们的目标是实现能自动检索出最确切描述查询文本本质的图像序列。为了更密集地填充图像样本，我们还利用了大量的照片流；每个照片流是由一个用户在一天之内拍摄的图片序列。为了学习文本和图像序列之间的相关关系，

我们采用的是利用潜变量【11,18,35】的基于结构排序的支持向量机的方法。

我们的研究可以广泛应用于网页服务应用程序中，尤其是旅游领域。例如，我们可以看到游客在 Tripadvisor 和 Yelp 上的纯文字评论通过自动从网上检索得到最形象化的图像序列。这个应用程序是很有意义的，因为一般用户可以理解关键概念并能够更容易更快的找到照片。除此以外，这些照片对新游客来说更有用。例如，那些从未去过迪斯尼乐园的用户可能不完全理解没有插图的关于景点的错误评论，但是我们的方法可以生成。

作为一个问题领域，我们专注于主题公园，特别是迪斯尼乐园，因为它很容易获得丰富的图像和文本数据。然而，我们的方法和问题构想是更广泛的，适用于拥有大量博客文章和图片的任何领域（更广泛地说，是任何图像和文本媒体）。一个具体的例子是讨论和评价博物馆、餐厅、城市或国家的旅游网站。在这种环境下，可以在评论中创建代表情绪波动的插图。我们这种方法是无监督的并且适用于任何一个数据可用的领域。

我们通过新收集的迪斯尼乐园的数据集来评估这种方法的图像检索性能，它包含超过了 10K 的博客文章与 120K 相关的图片，还有 6K 的超过 540K 的独特图像的照片流。我们用全面的实证研究来比较五个文本分科，三个文本描述，两个文字--图像嵌入方法和他们组合之间的图像序列的检索精度。使用潜变量的支持向量机方法可以有效地使用一个统一的方法来整合多个算法的输出。我们将用户在 Tripadvor 和 Yelp 上的评论进行可视化操作，并通

过亚马逊土耳其机器人研究的用户来评估。

1.1.相关成果

现在我们来讨论一些研究图像和自然语言文本之间关系的代表性成果。

根据图片/视频生成句子。这个成果的目标是对于一个给定的图像【8, 9, 17, 24, 34】能自动创建或检索出一个简洁的描述性的句子。这其中,【8】和【9】是与我们本次工作最相关的，因为他们的方法是直接利用大量的原始（可能未去噪）的在线数据，如【8】中的新闻文章（图片和文字）的多种模式的数据库，和【9】中 Flickr 带噪声标签、标题的描述的图片注释。在我们的工作中，除了 Flickr 的照片流，我们还开发了之前没有开发过的博客文章和消费者评论。此外，我们的工作着重于查询多个句子的图像序列检索的扩展问题。

图像和文本之间的映射关系。之前的工作也已经研究过句子和图像之间的映射关系或者从一个到另一个的检索。(例如【7,10,29,37】)。这项工作的主要重点是定义一个能够嵌入图片和句子的共同的语义空间。一些成功的想法包括:【7】中的三胞胎的对象、动作和场景,【10】中的核典型相关分析,【29】的依赖树形递归神经网络(DT-RNN),【37】中的抽象场景和条件随机场模型。我们工作新颖性的关键是,我们专注于多段落和图像序列之间的关系，而不是句子和图像之间的关系。

最近，多模式复发性神经网络【14,16,29】已经被广泛用于文本和图像之间的映射。我们潜变量的结构性支持向量机框架是具有吸引力的,因为它具有灵活性,这能使我们以一个统一的方式学习到组合的不同模型组件的权重，包括文本分类，文本描述符和文本-图像映射方法。

图像/视频检索结构化查询。这个方向的工作超越传统的关键字图像/视频检索，和地址结构化查询。一些著名的例子包括在自主驾驶【22】上下文中视频搜索一个句子,基于视觉词组的图像排序和检索【25】,多属性查询【27】,和图结构的对象查询【18】。【28】的工作提出了一种方法,合并三个不同的查询方式（如文本、素描和图像）为图像检索的语义空间。我们的工作独一无二在两个方面，第一，我们的查询结构是自然语言段落，第二，检索目标是图像序列而不是单独的图像。

最相关的成果之一是【13】，这也是实现了自动为故事配图的方法。然而，也存在重要差异。首先，给定一个查询通道,我们的目标是强调文章发展的检索图像序列,而不是类似的图像检索。第二，我们的方法是利用非结构化的网络博客和照片流,而不是专家创建的数据集。

1.2.贡献

据我们所知，这个工作是第一个处理多段自然语言查询指令中图像序列的排序和检索的。相较于先前的研究，我们的输入和输出扩展到更复杂的形式：段落而不是句子，和图像序列而不是单独的图像。

我们开发了一个建立在结构性排序的潜变量的支持向量机上的图像序列检索方法。我们的这种方法可以灵活地合并不同的文本和图像结构信息。

我们评估这种方法是有一个很大的非结构化迪斯尼数据集，包括 120K 相关图片的 10K 的博客文章，和 540K 图像的 6K 的照片流。经过定量评价和通过亚马逊土耳其机器人的用户研究，我们展示了我们的这种方法在实际用户书写的可视化自然语言文本方面是实用的。

2.问题公式化

我们有三种类型的输入数据。第一种输入是一系列游客的博客文章 $\mathcal{B}=\{B^1,\cdots,B^N\}$。我们假定每篇博客文章$B^n$由一个图像序列和相关文本组成 $B^n=\{(I_1^n,T_1^n),\cdots,(I_{|B^n|}^n,T_{|B^n|}^n)\}$。博客文章集$\mathcal{B}$是用作图像-文字平行语料库进行训练，从中我们可以实现联合图像-文字嵌入到一个共同的潜在空间。

第二种输入是大量游客照片流 $\mathcal{P}=\{P^1,\cdots,P^L\}$。我们定义了一个摄影师在一天内拍摄的一系列照片为一个照片流。照片流的主要用途是补充检索的图像样本。我们在同一潜在空间从博客数据使用转换方法实现照片流的嵌入，然后对于一个给定的查询文本，我们也可以从照片流中返回。

第三种是包含多个句子或段落的纯文本输入 $\mathcal{Q}$。每一篇文章$Q\in\mathcal{Q}$包括用户评论或没有图片的博客，我们用 $\mathcal{Q}$ 代表一组文本查询指令。

我们制定一个多段落查询指令中的图像序列的检索如下。给定一个查询文本$Q\in\mathcal{Q}$，我们对图像序列进行排序，$\mathcal{S}=\{(S^1,w^1),\cdots,(S^K,w^K)\}$，$S^k$是图像序列排序的第 k 个，$w^k$是排名得分，K 是检索集的数量。每个序列$S^k$ 由博客文章$\mathcal{B}$或照片流 $\mathcal{P}$组成。我们假设S^k是检索序列的大小，κ^k表示用户的输入，因为它是相当主观的个人偏好。例如，一些博客作者上传十多张照片作为一篇短的博客文章，然而其他人可能会更详细的在每篇博客里仅用几张图片。

2.1.文字-图像平行语料库

我们假设一篇博客中的每张照片在语义上与本博客的某些部分有相关关系。创建文本-图像并行训练语料库的挑战在于一篇博客中的文本通常是非结构性的，因此，贵文本和图像之间的规范化关系是未知先验的。获取注释器标签是一种选择，但是大型语料库是不可扩展的。因此，我们的方法是将一篇文章分割成语义连贯的几个文本片段，并且使每个片段的图像

之间有关联。然后，我们使用文本-图像块距离来确定图像文字的相关关系的不同程度（即一篇文档中，与图像距离较近的的文本段的关联程度高于距离较远的）。一旦得到了文本段和图像文本相关关系，每篇博客文章B^n都可以分解成一个图像序列和相关文本：$B^n = \{(I_1^n, T_1^n), \cdots, (I_{|B^n|}^n, T_{|B^n|}^n)\}$。

2.2.文本段

我们假设博客作者用图像、博客的文本段来在增强语义的方式。因此，文本分割的目的是把输入文本分为若干组相关的句子，并且，每一段都将与一张照片相关联。我们根据其句法结构和语义分布，在 MLP 文学【5,6,32】上应用了几种自动文本分类方法。从以下方法中产生的分段代表了一段话的中心内容，或者是文章中潜在的单个主题。我们实现了一个句法分割【1】和四种语义分割方法【2-5】。基于语义分割方法代表了每个句子的语义中心，并能聚合中心句来使每个分段保持其一致性。

（1）段落分割器。最简单的分段法之一是通过文章的语义结构，例如段落。通常每一段都带有独特的主题或事件，因此很可能带有相关图片。我们使用一种基于规则的正则表达式来检测段落划分的标准段落分割器（NLTK【23】）。

（2）潜在语义分析（LSA）。LSA 应用奇异值分解法（SVD）来获取句子的概念维度【19,30】。假设一段文章由多个概念或话题组成，每个都由文章中的一些术语所代表，基于 LSA 的方法可以递归地找到使每个主题（例如，最突出的主题边界）【32,5】的相似性值最大化的最具代表性的句子或一组句子（段）。

（3）LexRank 算法。LexRank 算法是根据文本中的词汇中心来检测出关键句子的。在一篇博客中，我们通过为每一句话创建一个顶点的方法来构建一个图形，并且使用 TFIDF 向量的内部句子余弦相似性方法来连接语义相似的句子构成边。通过随机漫步和特征向量中心的评估出的语义重要性，我们可以获得中心句子。对于每个句子图的重心，我们建立一个文本分段并且它的边界到两个句子图重心的距离相等。

（4-5）基于摘要的 LSA 和 LexRank 算法。LSA 和 LexRank 算法不仅可以进行分割，也可以进行摘要概述【6,15】。因此，使用这两种基本算法，我们可以进行分割和摘要概述，并将生成的每句概述都匹配上一张相关图片。（2）-（3）段代表了每次分割出多个句子，而（4）-（5）则只从每个文本段中选择语义最中心的一句话，并且可以删除不具有代表性的部分分段。

2.3.文本描述

执行文本分割后，我们从每一个文本段中提取特征。我们首先通过正常

化的分割段和删除的单词来对文本语料库进行预处理。我们采用三种标准文本描述符。

词包（BOW）。词包方法是一种简化的文本表示法，对于每个多句子输入文本，它可以忽视其复杂语义或语法【26】。

TF-IDF。TF-IDF 通过权重每个术语的出现频率和逆文档频率改善了 BOW 方法，从而能够识别给定文本中的特有的关键术语【1】。TF-IDF 可以有效地捕捉给定文本的特点，特别是长文本（例如多个段落）。因为 TF-IDF 的向量可以非常稀疏，我们可以通过只挑选出现频率最多的术语来把特征维数降到 20，000。

LDA 话题分配。LDA 模型可以代表每个作为潜在内容混合比例的给定文本，并通常解释为“话题”【3,4】。因此，LDA 模型可以用更简洁的维数来表达一个文本，，这已经在许多任务中被验证有效，甚至包括文本分类【21,31,33】。在我们的实验中，是一个 50K 主题，超过 2K 的语料库和 10K 博客文章的主题模型。

2.4.图像描述

我们使用密集特征提取与矢量量化的方法来进行图像的描述，这是一个近年来计算机视觉研究的标准方法。我们密集地提取出 HSV 色彩模型的 SIFT 方法并且分别在每张图的 4-8 为一步来把直方图的边缘安放在规则网格图内。然后，我们通过运用 K-mean 方法随机选择的描述符来形成 300 个视觉单词。最后，将最近的单词分配给网格的每个节点。作为图像或区域描述符，我们构建 L1 规范化空间金字塔直方图来计算每个视觉单词在三个级别的规则网格中的频率。我们通过连接这两个空间金字塔直方图的 HSV 色彩模型和 SIFT 方法来定义了图像描述符 v。

2.5.文本-图像嵌入

文本-图像的嵌入旨在获得图像及其相关文本之间的映射，并允许我们检索最接近给定文本的图像，反之亦然。我们实现了两种方法，其中包括一个使用 NCCA（规范化典型相关分析）【9】方法的参数，和一个使用简单的临近搜索的非参数方法。假设每个训练博客都被分割成文本和图像（例如 $B^n = \{(I_1^n, T_1^n), \cdots, (I_{|B^n|}^n, T_{|B^n|}^n)\}$），并且最后我们得到 M 对图像和文本。在之前的章节中，使用图像和文本描述符，现在我们分别用 x，y 代表每个图像和文本。然后，重复嵌入每个图像和文本的描述符，并且单独地细分方法。

（1）NCCA。我们提出分别用矩阵 $\mathbf{X} \in \mathbb{R}^{M \times d}$ 和 $\mathbf{Y} \in \mathbb{R}^{M \times D}$ 来代表 M 对图像和文本。然后文本-图像的嵌入目标是找到矩阵 $\mathbf{U} \in \mathbb{R}^{d \times c}$ 和 $\mathbf{V} \in \mathbb{R}^{D \times c}$ 来把图像和文本映射到一个由 $\mathbf{XU}$ 和 $\mathbf{YV}$ 的潜在空间的共同部分。CCA 的目的是找到这样的 U 和 V：

$$\max_{\mathbf{U},\mathbf{V}} \operatorname{tr}(\mathbf{U}^T\mathbf{X}^T\mathbf{Y}\mathbf{V}) \quad (1)$$
$$\text{s.t. } \mathbf{U}^T\mathbf{X}^T\mathbf{X}\mathbf{U}=\mathbf{I}, \mathbf{V}^T\mathbf{Y}^T\mathbf{Y}\mathbf{V}=\mathbf{I}.$$

CCA 优化是像解决【9】中的广义本征值问题一样的：

$$\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^T & \mathbf{C}_{yy} \end{pmatrix} \begin{pmatrix} \mathbf{z}_x \\ \mathbf{z}_y \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix} \begin{pmatrix} \mathbf{z}_x \\ \mathbf{z}_y \end{pmatrix}, \quad (2)$$

$$\mathbf{C}_{xx} = \mathbf{X}^T\mathbf{X}, \mathbf{C}_{xy} = \mathbf{X}^T\mathbf{Y}, \text{ and } \mathbf{C}_{yy} = \mathbf{Y}^T\mathbf{Y}$$

我们形成了每个 $\mathbf{z}_x$ 和 $\mathbf{z}_y$ 分别对应于投影

矩阵的 U 和 V 的特征向量。NCCA 提出了计算图像 x 和文本 y 之间相似性 $\sigma(\mathbf{x},\mathbf{y})$ ：

$$\frac{(\mathbf{x}\mathbf{U}\text{diag}(\lambda_{x1}^t,\ldots,\lambda_{xc}^t))(\mathbf{y}\mathbf{V}\text{diag}(\lambda_{y1}^t,\ldots,\lambda_{yc}^t))^T}{\|\mathbf{x}\mathbf{U}\text{diag}(\lambda_{x1}^t,\ldots,\lambda_{xc}^t)\|_2\|\mathbf{y}\mathbf{V}\text{diag}(\lambda_{y1}^t,\ldots,\lambda_{yc}^t)\|_2} \quad (3)$$

$\lambda_{x1},\ldots,\lambda_{xc}$ 是 c 特征值对应的特征向量 $\mathbf{z}_x$，而 t 是特征值的能力。我们在【9】中用 $c=96$ 和 $t=4$。使用 Eq【3】相似性度量，对于任意给定文本，我们可以检索出最接近的图像，反之亦然。

（2）KNN。KNN 是一种懒惰学习技术，这其中可以保留所有 M 训练对图像和文本。当 $\mathbf{x}' = \text{argmax}_{(\mathbf{x}',\mathbf{y}')\in\mathcal{T}} \cos(\mathbf{y},\mathbf{y}')$，图像 x 和文本 y 的相似性 $\sigma(\mathbf{x},\mathbf{y}) = \cos(\mathbf{x},\mathbf{x}')$ 。我们首先应从训练集中找到从 $\mathbf{y}'$ 到 $\mathbf{y}$ 的最接近文本，并计算相关于 $\mathbf{y}'$ 的从 $\mathbf{x}$ 到 $\mathbf{x}'$ 的余弦相似值。因此，$\sigma(\mathbf{x},\mathbf{y})$ 和 $\sigma(\mathbf{y},\mathbf{x})$ 的值并不相同，因为前者是从一个图像空间计算出的，而后者是文本空间。

在第 4 节实验中，我们将比较这两个嵌入方法的检索性能。

3.检索模型

我们设计了基于潜变量的结构化支持向量机的排序和检索的方法（例如【12, 35】）。判别函数被定义为一个实值函数 $\mathcal{F}(Q,\mathcal{S}):\mathcal{Q}\times\mathcal{S}\rightarrow\mathbb{R}^+$，能够衡量一个查询的兼容性文本 $Q\in\mathcal{Q}$ 和图像序列 $S\in\mathcal{S}$。

自然，从文本映射到一个图像序列取决于如何分割给定的文档。我们称之为文本分割，它定义了函数 $\mathcal{G}(Q,\kappa,H)$，把查询段 Q 划分成一系列 κ 段 $H=\{h_1,\cdots,h_\kappa\}$，因此，每个文本段 h_i 都是总体连贯的并映射到一个单独的图像上。因此，图像序列 $S=\{s_1,\cdots,s_\kappa\}$ 有相同的长度 $\bar{H}$，其中每个 s_i 都对应于 h_i。在实践中，每个 h_i 都可以是单个或者多个句子。

我们将文本段的输出作为一个潜变量，因为它的正确答案在训练和测试阶段是不可用的。如果，我们用 $H\in\mathcal{H}$ 表示一个段落分割的实例，查询文本 Q 的检索目的是找到最优图像序列 S^*：

$$S^* = \underset{(S,H)\in\mathcal{S}\times\mathcal{H}}{\text{argmax}}\ \mathcal{F}(Q,S,H) = \underset{(S,H)\in\mathcal{S}\times\mathcal{H}}{\text{argmax}}\ \mathbf{w}\cdot\Psi(Q,S,H) \quad (4)$$

通常的判别函数是线性特征向量 $\Psi(Q,S,H)$，它描述了查询输入 Q 和图像序列 S 之间的关系，还有作为潜变量的分割实例 $\bar{H}$ 。

3.1.特征空间

我们将特征向量分为两个部分：

$$\mathbf{w}\cdot\Psi(Q,S,H) = \boldsymbol{\alpha}\cdot\Phi(Q,S,H) + \boldsymbol{\beta}\cdot\Pi(Q,S,H). \quad (5)$$

第一个内容 $\Phi(Q,S,H)$ 是包括了一组描述每一对一对一关系 (s_i,h_i) 的特征，而 $\Pi(Q,S,H)$ 则由 S 和 H 作为一组相关的特征向量组成。

第一个特征集$\Phi(Q,S,H)$，测量了文本和图像之间一对一的兼容性，连接了 2.4 节中的两个图像特征和 2.3 节中三个文本特征所有组合的平均相似特性。因此$\Phi(H,S)\in\mathbb{R}^6$被定义为：

$$\Phi(H,S)=\frac{1}{\kappa}\left[\sum_{i-1}^{\kappa}\sigma(\mathbf{x}_i^1,\mathbf{y}_i^1)\cdots\sum_{i-1}^{\kappa}\sigma(\mathbf{x}_i^2,\mathbf{y}_i^3)\right]^T, \quad (6)$$

$\mathbf{x}_i^1$ 和$\mathbf{y}_i^1$分别是第一类型的图像h_i和文本描述符s_i。求图像文字的相似性$\sigma(\mathbf{x},\mathbf{y})$，我们使用 2.5 节中定义的两种方法之一（如 NCCA 和 KNN）. 第二个特性集$\Pi(Q,S,H)$描述了文本段H和图像S作为一个整体的兼容性。我们使用两个流行的相似性方法：一个乱序的法规，Hausdorff 相似，和一个有序的动态时间规划（DTW）相似。我们同时考虑乱序和有序的规则，因为在不同的博客中图像-文字的顺序关系的可能并不总是一直匹配一致的。最终的特征维度为$\Pi(Q,S,H)\in\mathbb{R}^{12}$（即 2 度量单位*3 文本*2 图像特征）。

3.2. 学习

为了学习这种模型(如计算 Eq. (5) 参数向量 w),我们使用博客文章的数据作为训练数据。符号的轻微滥用，我们用$\mathcal{B}_t=\{(Q^{(n)},S^{(n)})|n=1,\cdots,N\}$表示我们的训练数据，$Q^{(n)}$ 和 $S^{(n)}$是训练博客n的文本和图像序列。结构化的潜在支持向量机的学习模型是：

$$\min_{\mathbf{w},\xi\geq 0}\ \frac{1}{2}\|\mathbf{w}\|^2+\frac{C}{N}\sum_{n-1}^{N}\xi_n \quad (7)$$

$$\text{s.t. }\mathbf{w}\cdot\left(\Psi(Q^{(n)},S^{(n)})-\Psi(Q^{(n)},S)\right)\geq\Delta(S^{(n)},S)-\xi_i, \quad (8)$$
$$\forall n,\forall S\in\mathcal{S}\backslash S^{(n)}$$

ξ_n 是一个减弱变量， C 是一个正则化参数。损失函数被定义为 ：

$$\Delta(S^{(n)},S)=1-\frac{1}{|S^{(n)}|}\sum_{s-1}^{|S^{(n)}|}\sigma(s_i^{(n)},s_i)$$

这是位置变量， $S^{(n)}$和S之间的平均距离。我们令$S^{(n)}$等于S的长度。

请注意，在训练和测试时间中，我们没有获得正确的分割文本$Q^{(n)}$。因此，我们将每一个文本的分割段都成为一个潜变量。更具体地说,在 2.2 节中,我们描述了关于Q和κ的函数的五种不同的分割方法。作为每一个训练数据的潜变量,我们引入了$\mathbf{h}^{(n)}\in\mathcal{R}^D$， D 是可能的分割方法的数量。(如我们设置 D = 5)。因此，$\mathbf{h}^{(n)}$是一个只有一个非零元素的二进制向量并表示了分割输出被应用到训练文档 n 中。由于有潜变量，Eq. （8）可以被定义为：

$$\mathbf{w}_d\cdot\max_{\mathbf{h}_d^{(n)}\in\{1,\cdots,D\}}\left(\Psi(Q^{(n)},S^{(n)},\mathbf{h}_d^{(n)})-\Psi(Q^{(n)},S,\mathbf{h}_d^{(n)})\right)$$
$$\geq\Delta(S^{(n)},S)-\xi_i. \quad \forall n,\forall S\in\mathcal{S}\backslash S^{(n)}. \quad (9)$$

现在我们有用 d 种方法从博客文章中获得的 D 种不同的$\mathbf{w}_d$集。

在 Eq. (9)，$\mathcal{S}$可以是任何可能的图像序列 κ 的大小。由于$S\in\mathcal{S}$可以是无穷的，我们现实负数$\mathcal{S}$的产生如下：对于每个训练博客n，有一个固定的η，我们通过对$S^{(n)}$随机应用两种方法产生的负数$\mathcal{S}$，这是 1997 年实现的用其他的图片来代替某些$S^{(n)}$。我们在实验中设置$\eta=50$。

优化。我们对潜在结构的支持向量机使用交替优化方法（例如【18,35】）。总之，我们替换了支持向量机优化的变量$\{\mathbf{w}_d\}_{d-1}^{D}$，并且划分了其标签 h。我

们对每一篇训练博客第一次进行随机初始化 $\mathbf{h}^{(n)}$并应用相应的分割方法 d。然后，我们重复以下两个步骤，直到收敛或者预定的迭代次数。

对每一个含有 $\mathbf{h}^{(n)}$的训练数据 n 都有一个固定的分割方法，我们解决 Eq.(7)中标准的结构化的支持向量机来获得 $\{\mathbf{w}_d\}$。我们使用【12】中提出的边缘尺度的 n-slack 算法。

对所有训练博客$n \in \{1,\cdots,N\}$通过 $\mathbf{h}^{(n)} = \text{argmax}_d \mathbf{w}_d \cdot \Psi(Q^{(n)}, S^{(n)}, H)$ 来更新分割方法，并固定支持向量机参数。

3.3. 推理

在测试中，我们给定了一组已知参数$\{\mathbf{w}_d\}_{d-1}^{D}$，一个查询指令文本$Q$，训练博客 $\mathcal{B}_t$和从中选出最匹配图像S^*的照片流数据库 $\mathcal{P}_t$。检索执行如下，假设候选图像序列$\mathcal{S}_{cand}$已经给定。然后，对于每个 $S \in \mathcal{S}_{cand}$，我们通过 $\max_d \mathcal{F}(Q,S,H) = \max_d \mathbf{w}_d \cdot \Psi(Q,S,H)$ 计算分数 S。也就是说，我们对应用了 D 种不同的分割方法，通过在 D 种不同分割与它们相对应 $\mathbf{w}_d$的输出中找到最大的分数 S。最终，我们可以根据分数对 $S \in \mathcal{S}$ 排序。

然而，这种情况的一个主要困难是在$\mathcal{S}_{cand}$中有许多候选项，所以，我们使用一个近似策略。一旦查询指令文本被分割成$H = \{h_1,\cdots,h_\kappa\}$，我们首先找到每一个 h_i的最邻近图像，并约束$S \in \mathcal{S}$从其中产生。

在这种约束下，$\mathcal{S}_{cand}$的大小是 $K_h \times \kappa$。进一步地提高S^*的搜索速度，我们可以使用贪心算法；例如，一旦我们为h_1选择了一个图像，我们为h_2选择下一个图像就要最大化 $\mathbf{w}_d \cdot \Psi(Q,S,H)$。贪心算法已广泛用于结构化支持向量机的子集选择问题的应用程序上【20,36】。

3.4 实验

我们第一次进行全面的图像序列的检索任务的实验，来比较不同文本特征，分割方法，嵌入方式和它们组合的成果。然后，我们证明了我们的方式具有将 TRIPADVISOR 和 YELP 上纯文本的评论进行可视化的能力。我们用亚马逊土耳其机器人所获取的一般用户的偏好来进行用户研究。

4.1. 照片流，博客和评论的数据集

我们抓取了加利福尼亚州的两个公园的图像和文本数据：迪士尼加州冒险和迪斯尼乐园公园。

照片流数据。我们通过从 Flicker 的 6026 照片流上查询有关于迪斯尼乐园的关键字收集了 542,217 张独特的照片。我们只考虑包含超过 30 张图片并去除掉与公园无关噪声的照片流。

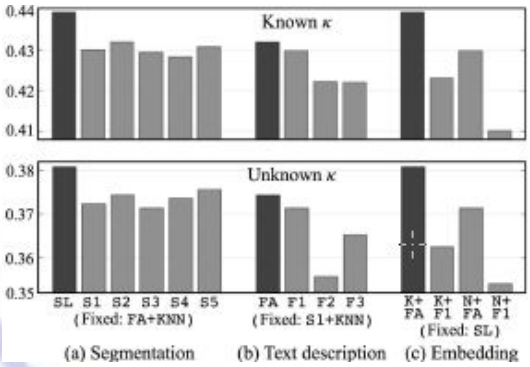
博客数据。我们第一次从三个流行发表博客的网站 BLOGSPOT, WORDPRESS, 和 TYPEPAD 上，通过从谷歌搜索改变查询条件来抓取了 53,091 篇独特的博

客文章还有 128,563 张相关图片。然后，博客是通过迪斯尼专家手工分成三类：游记，迪斯尼和垃圾。游记标签表明我们感兴趣的博客文章是用多个图片来描述迪斯尼乐园的故事和事件。迪斯尼标签是被用于标记那些与迪斯尼相关但不是游记，如迪斯尼乐园的历史，电影或者商品的文章。我们只使用了大小为 10,075 篇文章和 121,252 张相关图片的游记标签的博客文章。

TripAdvisor 和 Yelp 数据集。 TripAdvisor 和 Yelp 在它们的网站上为游客提供特定的评论场所。我们手动地从每个网站（总共 200 个）上挑选出关于加州迪斯尼州的冒险和迪斯尼乐园的 100 篇评论。我们挑选出的评论既不太长也不太短。我们用游客的评论数据是为了评估和证明从博客数据中获得的图像-文本相关关系可以被灵活应用于相同领域的其他类型的文本输入。

4.2. 图像序列检索的结果

对于定量评估，我们随机选择了80%的博客文章作为一个训练集，剩下的作为另一个训练集。为了测试每一篇博客文章，我们使用文本的部分作为查询文本Q，并且图像序列作为标本S_G。每个算法的目标是从训练博客或照片流$\mathcal{P}$中获取最佳图像序列。由于训练和测试数据是不相交的，所以每个算法最多只能检索相似（但不相同）的图像。我们在两个不同的设置下进行实验:在给定算法下，有或没有出现在原博客中的标准图像序列的大小。我们测试了 10 个不同的训练集和测试分区。为了评测性能，我们需要对一个查询文本Q定义$S=\{s_1,\cdots,s_K\}$如何接近$S_G=\{s_{G1},\cdots,s_{G|S_G|}\}$。因为检索序列只能和标准样本相似，并且可能没有相同数量的照片，我们定义如下的基于相似性的 Jaccard 指数作为一个评价指标。我们首先用第二节中的规范化描述符L_1-代表图像。因为S和S_G可能是长度不相等的两个向量序列，我们用动态时间规划（DTW）算法来对齐S和S_G，这种方法能够发现它俩之间的一组对应关系$\mathcal{C}$。然后，我们定义基于相似性



(a)分割　(b)文本描述　(c)嵌入

图 2 比较 Eq.(10)中的基于相似性 Jaccard 指数方法的图像检索精度：（a）中分割方法不同，（b）中文本表示方法不同，（c）中嵌入方法不同，其中标准样本的图像序列大小已经给定（上），并且算法是未知的（下）。

的 Jaccard 指数为：

$$J(S,S_G)=\frac{\sum_{(s_i,s_j)\in\mathcal{C}}\sigma(s_i,s_j)}{\max(|S|,|S_G|)} \qquad (10)$$

$\sigma(s_i,s_j)$是余弦相似性，注意在外形和长度方面，$J(S,S_G)$比S和S_G都更相似。

因为在之前的文学作品中，多段落查询文本的图像序列的检索问题尚未解决，我们只能对我们提出的方法的不同组合进行综合比较。具体来说，作为基准，我们变换了 5 种文本分割方法（语法（S1），LSA（S2），基于 LSA 的简单摘要（S3），LexRank（S4）和基于 LexRank 多的简单摘要（S5）），三种文本表示方法(词包(F1)，TF-IDF（F2）和 LDA 主题分布（F3）），和两种文本-图像嵌入方法（（KNN）和（NCCA））。这些基准是经过我们潜在的结构化支持向量机模型对每一篇输入文章（SL）动态分配最好的分割方法和联合优化功能权重（FA）后得到的。

比较文本分割方法。我们测试了相对于固定分割方法我们的潜在模型是否提高了检索性能。为了这个实验，我们固定嵌入方法为（KNN），文本描述方法为（FA）。

图 2 显示了我们的潜在模型（SL）可以给每个输入博客的文章分配最好的分割方法，超越了所有使用一种固定分割的基线方法。由于每个博客的写作风格不同，该结果的直观感受就是每一篇输入文章要发现和对其应用最佳分割算法以改善性能。显然，当基准图像序列大小(κ)已指定时算法能更好的执行，检索出偏离基准图像最小的图像。需要注意的是，即使使用 Jaccard 指数测量增加了小数字的精确度，却表示了显著质量的提高，这一点将在 4.3 节中得到印证。

vs. (FA+NCCA)	# Votes	5	4	3	2	1	0
60.6% (303/500)	# Samples	15	26	21	26	9	3
vs. (F1+KNN)	# Votes	5	4	3	2	1	0
59.0% (295/500)	# Samples	16	19	29	19	14	3
vs. (S2+KNN)	# Votes	5	4	3	2	1	0
57.4% (287/500)	# Samples	11	21	32	18	16	2

表 1 在我们的方法和三个基线之间，通过 AMT 方法对消费者的评论进行可视化配对偏好测试的结果。数据表示我们可视化的百分比高于给定句子的基线。

文本特征比较。当其中一个文字描述符是固定的时，我们测试了与基线相比，我们的模型是否改善了检索性能。针对这个实验，我们固定分割和嵌入方法到(S1)+(KNN)。图 2.(b)显示，聚合的功能集（FA）优于单一表现最好的功能集。在这些文本特征中，TF-IDF (F2)表现最好。结果清楚地表明，从不同的表示获得的文本信息是互补的。并且我们的结构化支持向量机模型成功地调整了每对图像和文本特征（FA）的重要性。

文字到图像嵌入方法的比较。图 2.(C)对比了当应用两个不同的特征描述（FA）和（F2）时两种嵌入方法的性能。我们固定分割方法（SL）。在我们的实验中，非参数方法（KNN）得到了比参数（NCAA）更好的图像和文本的映射。我们观察到了最好的一对嵌入方法和文本表示方法是(KNN)+(FA)。

总之，我们的方法始终优于基准的 2.5%至 8%。精度定量提高的表现是中等的，主要是因为 Eq.(10)基于相似性的指标可以正确的编码除了抑制感知差异（类似于 BLEU 分数）。

定性结果。图 3 显示了图像序列检索

的例子。我们用标准样本比较了我们的算法和基线分别得到的结果。我们的算法说明文档的主旋律是通过在景点（例如米奇卡通城），活动（例如游行）和地点（例如餐厅）方面检索相关景点。

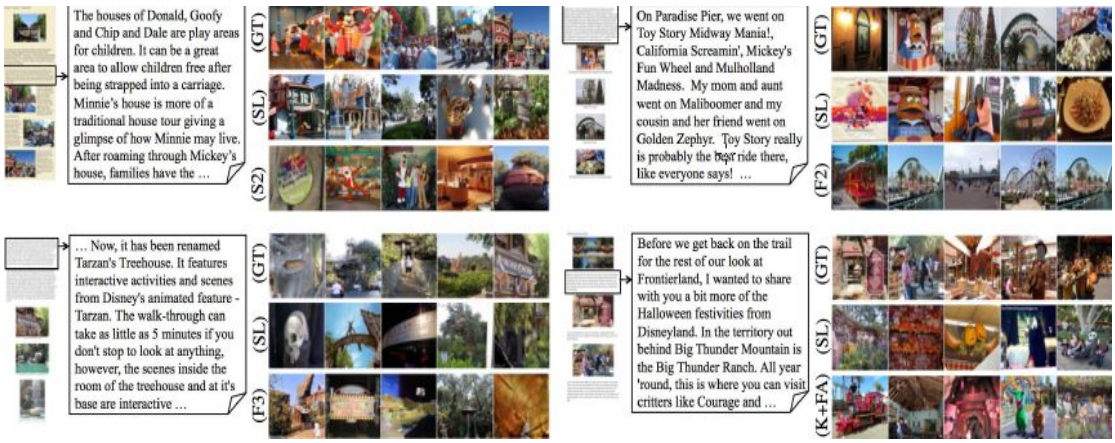


图 3　定性比较图像序列的检索。左图，我们展示了规模原创博文和部分查询文本，右图，我们展示了标准样本图像序列（GT），预测算法（SL），和从上到下的一条基线。更相似的标准样本的预测序列，对于给定查询文本更准确的检索算法。注意实际博客和图像序列更长，我们只是出于演示的目的显示部分片段。

图四　消费者在 TRIPADVISOR 和 YELP 上的可视化的评论例子。我们用加粗字体展示文档中与可视化检索图像相关的部分。对于每一个文本查询，我们展示前三个图像序列。注意实际查询评论和图像序列都更长，我们只是出于演示的目的显示部分片段。

4.3 消费者评论的可视化

我们评估我们的算法对一般用户的评论可视化的能力。我们从 TRIPADVISOR 和 YELP 数据集中通过选择 100 篇有故事情节的评论组成一组查询文本，这些评论主要描述了顺序访问流。由于查询评论是纯文本的，并且没有标准样本图像序列，我们通过亚马逊土耳其机器人（AMT）来说会用基于资源的人群评估。我们首先使用博客数据训练方法和作为基线。然后展示每一个查询文本和一对通过算法预测和随机顺序之一的基线所得的图像序列，并询问土耳其人来选择对

文章可视化最好的那一个。我们的每个查询都是从 5 个不同的土耳其人那里得到的回答。

表 1 显示了成对的 AMT 偏好的结果测试。我们比较了每个使用不同的文本特征,分割和嵌入方法的三个基线。虽然问题在本质上是高度主观的，并且有各种各样的同样好的答案，但是我们的结论是通过大量土耳其人的喜好而得出的。

图 4 展示了 TRIPADVISOR 和 YELP 的评论中前三个的图像序列。尽管实际查询评论和图像序列都更长，我们只是出于演示的目的显示部分片段。我们强调了我们通过我们的算法实现文本可视化的条款。迪斯尼乐园提供了一个多样化的且文本和图像高度共存的娱乐活动，事件和景点的数据集。我们的方法可以帮助构建它们之间的交叉引用，具有广阔的应用前景和可以应用到网络应用程序中。

5. 总结

我们提出了一个用于多段落查询文本的图像序列检索的方法。使用网上的博客文章和照片流，我们构建了一个研究图像-文本关系的并行语料库。我们确定了一个潜在的结构化支持向量机方法来研究他们之间的语义关系，并通过 AMT 方法中大量的基线和用户研究得到了一个综合性的评估。

References

[1] A. Aizawa. An Information-Theoretic Perspective of TF-IDF Measures. *Info. Proc. Manag.*, 39(1):45–65, 2003. 4

[2] D. Beeferman, A. Berger, and J. Lafferty. Statistical Models for Text Segmentation. *Mach. Learn.*, 34(1-3):177–210, 1999. 3

[3] D. M. Blei and J. D. Lafferty. Dynamic Topic Models. In *ICML*, pages 113–120, 2006. 4

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 3:993–1022, 2003. 4

[5] F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent Semantic Analysis for Text Segmentation. In *EMNLP*, 2001. 3

[6] G. Erkan and D. R. Radev. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *JAIR*, 22(1):457–479, 2004. 3, 4

[7] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences from Images. In *ECCV*, 2010. 1, 2

[8] Y. Feng and M. Lapata. Automatic Caption Generation for News Images. *IEEE PAMI*, 35(4):797–812, 2013. 1, 2

[9] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. In *ECCV*, 2014. 1, 2, 4

[10] M. Hodosh, P. Young, and J. Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR*, 47:853–899, 2013. 1, 2

[11] T. Joachims. Training Linear SVMs in Linear Time. In *KDD*, 2006. 2

[12] T. Joachims, T. Finley, and C. N. J. Yu. Cutting-plane Training of Structural SVMs. *Mach. Learn.*, 77:27–59, 2009. 5, 6

[13] D. Joshi, J. Z. Wang, and J. Li. The Story Picturing Engine: A System for Automatic Text Illustration. *ACM TOMM*, 2(1):68–89, 2006. 2

[14] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *NIPS*, 2014. 2

[15] K. Kireyev. Using Latent Semantic Analysis for Extractive Summarization. In *TAC*, 2008. 4

[16] R. Kiros, R. Salakhutdinov, and R. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In *TACL*, 2015. 2

[17] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby Talk: Understanding and Generating Image Descriptions. In *CVPR*, 2011. 1, 2

[18] T. Lan, W. Yang, Y. Wang, and G. Mori. Image Retrieval with Structured Object Queries Using Latent Ranking SVM. In *ECCV*, 2012. 2, 6

[19] T. A. Letsche and M. W. Berry. Large-scale Information Retrieval with Latent Semantic Indexing. *Information Sciences*, 100(1–4):105–137, 1997. 3

[20] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu. Enhancing Diversity, Coverage and Balance for Summarization through Structure Learning. In *WWW*, 2009. 6

[21] C. Lin and Y. He. Joint Sentiment/Topic Model for Sentiment Analysis. In *CIKM*, pages 375–384, 2009. 4

[22] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *CVPR*, 2014. 2

[23] E. Loper and S. Bird. NLTK: The Natural Language Toolkit. In *ETMTNLP*, 2002. 3

[24] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*, 2011. 1, 2

[25] A. Sadeghi and A. Farhadi. Recognition Using Visual Phrases. In *CVPR*, 2011. 2

[26] G. Salton, E. A. Fox, and H. Wu. Extended Boolean Information Retrieval. *CACM*, 26(11):1022–1036, 1983. 4

[27] B. Siddiquie, R. S. Feris, and L. S. Davis. Image Ranking and Retrieval based on Multi-Attribute Queries. In *CVPR*, 2011. 2

[28] B. Siddiquie, B. White, A. Sharma, and L. S. Davis. Multi-Modal Image Retrieval for Complex Queries using Small Codes. In *ICMR*, 2014. 2

[29] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded Compositional Semantics for Finding and Describing Images with Sentences. In *TACL*, 2013. 1, 2

[30] J. Steinberger and K. Jezek. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. In *ISIM*, 2004. 3

[31] S. Tasci and T. Gungor. LDA-based Keyword Selection in Text Categorization. In *ISCIS*, 2009. 4

[32] Y. Wang and J. Ma. A Comprehensive Method for Text Summarization Based on Latent Semantic Analysis. *NLPCC*, 400:394–401, 2013. 3

[33] Z. Wang and X. Qian. Text Categorization Based on LDA and SVM. In *CSSE*, pages 674–677, 2008. 4

[34] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2T: Image Parsing to Text Description. *IEEE Proc*, 98(8):1485–1508, 2010. 2

[35] C.-N. J. Yu and T. Joachims. Learning Structural SVMs with Latent Variables. In *ICML*, 2009. 2, 5, 6

[36] Y. Yue and T. Joachims. Predicting Diverse Subsets Using Structural SVMs. In *ICML*, 2008. 6

[37] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the Visual Interpretation of Sentences. In *ICCV*, 2013. 1, 2