计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700006

基于聚类网络的文本-视频特征学习

张 衡[1] 马明栋[2] 王得玉[2]
1 南京邮电大学通信与信息工程学院 南京 210003
2 南京邮电大学地理与生物信息学院 南京 210003
(1217012230@njupt.edu.cn)

摘 要 综合理解视频内容和文本语义在很多领域都有着广泛的研究。早期的研究主要是将文本-视频映射到一个公共向量空间,然而这种方法所面临的一个问题是大规模文本-视频数据集不足。由于视频数据存在较大的信息冗余,直接通过3D网络提取整个视频特征会使网络参数较多且实时性较差,不利于执行视频任务。为了解决上述问题,文中通过良好的聚类网络聚合视频局部特征,并可以同时利用图像和视频数据训练网络模型,有效地解决了视频模态缺失问题,同时对比了人脸模态对召回任务的影响。在聚类网络中加入了注意力机制,使得网络更加关注与文本语义强相关的模态,从而提高了文本-视频的相似度值,更有利于提高模型的准确率。实验数据表明,基于聚类网络的文本-视频特征学习可以很好地将文本-视频映射到一个公共向量空间,使具有相近语义的文本和视频距离较近,而不相近的文本和视频距离较远。在MPII和MSR-VTT数据集上,基于文本-视频召回任务来测评模型的性能,相比其他模型,所提模型在两个数据集上进行精度均有提升。实验数据表明,基于聚类网络的文本-特征学习可以很好地将文本-视频映射到一个公共向量空间,从而用于文本-视频召回任务。

关键词:召回模型;模态融合;聚类网络;视频理解
中图法分类号 TP391

Text-Video Feature Learning Based on Clustering Network

ZHANG Heng[1], MA Ming-dong[2] and WANG De-yu[2]
1 College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2 College of Geographical and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract Comprehensive understanding of video content and text semantics has been widely researched in many fields. The early research is mainly to map text-video to a common vector space. However, one of the problems faced by this method is the lack of a large-scale text-video datasets. Because of the large information redundancy of the video data, extracting the whole video feature directly through 3D network will lead to more network parameters and poor real-time performance, which is not conducive to video tasks. In order to solve the above problems, this paper proposes that the local characteristics of video can be aggregated by good clustering network, and the network model can be trained by image and video datasets at the same time to effectively solve the problem of video modal missing. At the meantime, the influence of face mode on recall task is compared. The attention mechanism is added to the clustering network, which makes the network pay more attention to the modes strongly related to the text semantics, so as to improve the similarity value of the text-video and improve the accuracy of the model. The experimental result shows that text-video feature learning based on clustering network can map text-video to a common vector space, so that text and video with similar semantics are close to each other, text and video with different distances are far away. In this paper, the performance of the text-video recall task evaluation model based on MPII and MSR-VTT datasets is improved compared with other models. From the experimental result, it is fully proved that the text-feature learning based on clustering network can map the text-video to a common vector space, which can be used in the text-video recall task.

Keywords Recall model, Modal fusion, Clustering network, Video understanding

到稿日期:2019-06-30 返修日期:2019-09-22 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:江苏省自然科学基金青年基金(BK20140868)
This work was supported by the Youth Fund of Jiangsu Natural Science Foundation (BK20140868).
通信作者:马明栋(mmdbs@126.com)

1 引言

视频内容理解在动作行为识别、视频检索、视频描述和安防等诸多领域有着广泛的研究。综合理解视频内容和文本语义，对人机交互的发展有着很大的推动作用。当前基于文本-视频的研究主要包括视觉问答[1-2]、视频描述、视频召回[3-6]等。这些研究主要是将文本-视频映射到一个公共向量空间，使具有相近语义的文本和视频距离较近，而语义不同的文本和视频距离较远，这类方法被广泛用于文本-视频模型[4-6]。

由于学习视频特征需要大量的训练样本[7-8]，而带有正确标签和视频描述的视频数据集较少，因此模型学习面临着较大的困难。目前常用的包含视频描述的数据集有 MPII 和 MSR-VTT[9-10]。采用迁移学习方式能使模型取得较好的效果，即将模型在 COCO，ImageNet，Visual Genome[11-13]等带有正确标签和图片描述的图像数据集中进行预训练，再利用视频数据微调网络参数。由于视频数据中包含运动信息和声音信息，图片中只包含视觉信息和人脸信息，因此迁移学习的效果不佳；同时，视频数据具有较大的冗余性，直接利用 3D 网络提取视频特征使得网络参数较多。

基于上述问题，本文提出了一种新的网络模型，该模型可以同时利用图像和视频数据训练网络，有效解决了视频模态缺失问题。本文的模型主要是学习将文本和视频映射到一个公共向量空间，然后利用距离来衡量两者之间的相似度。同时，针对不同的视频模态学习对应的模态权重，类似于注意力机制，以提升模型的精度。在训练过程中，当样本为图片数据时，认为样本缺失运动和音频信息。为了降低视频信息的冗余，本文首先从视频中抽取图像帧，然后通过良好的聚类网络来得到视频特征，从而增强模型的实时性。在 MPII 和 MSR-VTT 数据集上进行测试，结果表明，本文模型在文本-视频召回任务上取得了很好的结果。

本文模型具有很好的扩展性，支持各种模态的扩充，包括人脸。为了验证模型的可扩展性和探究人脸模态对视频召回效果的影响，将在视频中提取的人脸作为特征输入网络模型。

本文的主要贡献在于：1）提出了一个学习文本-视频特征模型，解决了视频模态缺失问题；2）通过 NeXtVLAD 聚类网络将局部特征聚合成视频特征，可以很好地提升模型的准确率；3）通过对比实验验证了所提模型能取得很好的结果。

2 相关工作

2.1 视频中的各种模态特征

本文使用以谷歌新闻为样本训练得到的 word2vec[14] 词嵌入模型来提取文本中的单词向量，为了得到句子特征，使用 NeXtVLAD[15] 网络将词向量聚合成句子向量。对于视频数据，每秒抽取一帧，取前 25 帧，然后将每帧图像缩放到 300 像素。本文使用不同的网络模型来提取不同的模态特征：使用在 ImageNet 数据集上预训练的 ResNet-152[16] 网络来提取视觉特征，从最后一层全局池化层提取 2048 维向量作为图像特征；使用在 Kinetics 数据集上预训练的 I3D[17] 网络提取运动特征，从最后一层全局池化层提取 1 024 维向量作为运动特征；通过语音卷积网络提取音频特征[18]，得到 128 维的特征向量；对于人脸特征，首先使用 DLib 框架检测和对齐人脸，用 ResNet 网络提取 128 维的特征向量。每种模态都是通过 NeXtVLAD 网络聚合得到整体的特征向量。

2.2 模态融合的方式

视频中主要包含视觉、音频和运动 3 种模态信息。目前大多数关于视频处理的算法主要使用视觉特征，对音频和运动特征使用得较少。而音频和运动同时包含了大量与视频相关的信息，因此将几种信息融合能更好地表征视频内容。相比图像而言，由于视频任务需要更多的训练样本，而大多数视频存在版权问题，导致公开的数据集较少，因此需要在一定的数据规模下提升模型精度。

由于视频包含多种模态信息，因此融合这些模态信息是提升模型精度的关键。目前，模态融合方式主要分为前融合和后融合两种方式，针对不同的数据集和不同的学习任务，两种方式获得的结果存在差异。图 1 所示为前融合方式，即先将各个模态的特征融合成一个新的向量，再通过线性变换将该向量映射到一个新的向量空间，使新特征向量能够很好地表征原始特征。

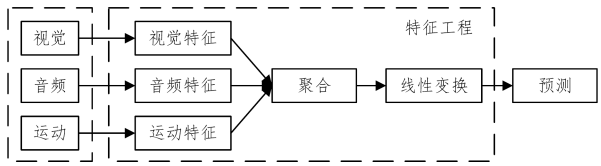


图 1 前向融合结构

Fig. 1 Forward fusion structure

图 2 所示为后融合方式，即先将每种特征映射到新的向量空间，然后进行融合。特征融合的基本操作包括点积、级联、平均等。

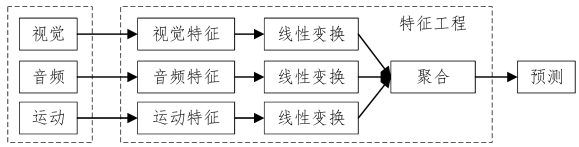


图 2 后向融合结构

Fig. 2 Backward fusion structure

3 基于聚类网络的文本-视频特征学习

3.1 整体网络框架

本文模型的目标是将文本-视频映射到一个新的向量空间，然后计算两者之间的相似度。假定 X 为文本，Y 为视频，当两者语义相近时，本文模型可以学习到映射函数 $f(X)$ 和 $g(Y)$，使得两者的相似度值 $s=\langle f(X), g(Y)\rangle$ 较大。本文假定输入的视频包含 N 种模态，用 $\{I_i\}_{i\in 1,\cdots,N}$ 表示。

图 3 给出了本文模型的框架。首先将每种模态 I_i 输入聚类网络 h_i 得到聚合向量，然后通过门控函数 g_i 得到最后的输出向量。相同文本所对应的每个词向量通过聚类网络得到句子向量，然后通过门控函数 f_i 得到最后的输出向量。每种模态的权重 $w_i(\boldsymbol{X})$ 通过句子向量 $\boldsymbol{X}$ 学习得到，最后计算文本-视频的相似度值。

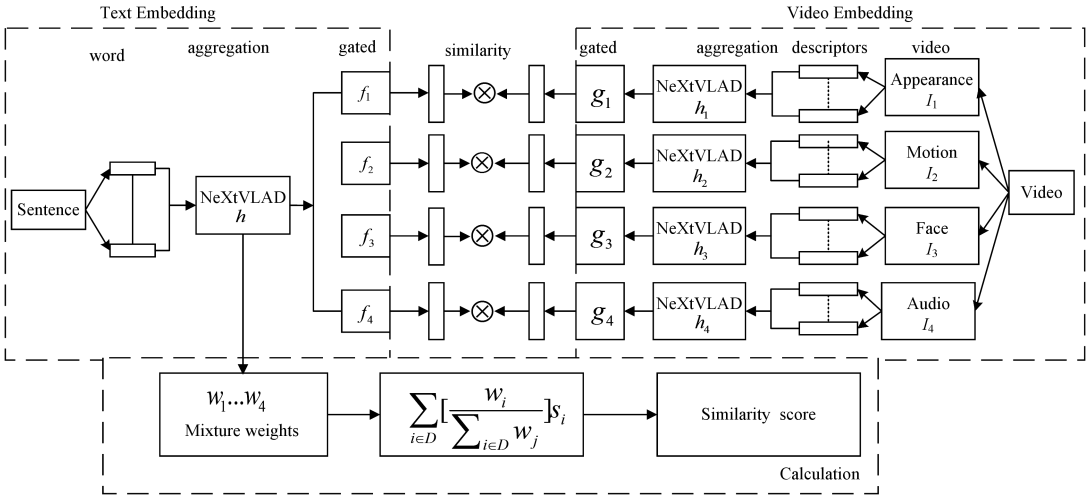


图 3　整体网络框架

Fig. 3　Overall network framework

3.2　聚类网络

本文通过 NeXtVLAD 网络将模态 I_i 的局部特征聚合成整体特征。NeXtVLAD 的网络结构，如图 4 所示。

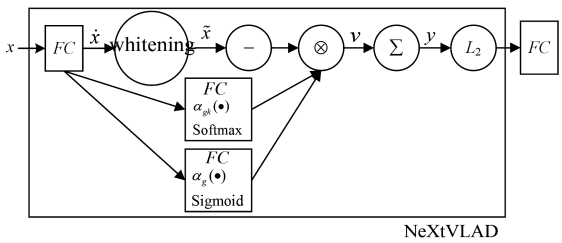


图 4　NeXtVLAD 的结构

Fig. 4　NeXtVLAD structure

在 NeXtVLAD 聚类网络中，输入特征向量 $\boldsymbol{x}_i$ 通过全连接层扩展为 $\dot{x}_i$，其向量维度为 λN，在实验中 λ 通常设定为 2。然后，通过分组操作将 $\dot{x}_i$ 的维度由 $(M,\lambda N)$ 变为 $(M,G,\frac{\lambda N}{G})$，其中 G 为分组数。这个过程等价于将 $\dot{x}_i$ 分解为 G 个低维的特征向量 $\{\tilde{x}_i^g \mid g\in\{1,\cdots,G\}\}$，即每个图像特征可以表示为：

$$v_{ijk}^g=\alpha_g(\dot{x}_i)\alpha_{gk}(\dot{x}_i)(\tilde{x}_{ij}^g-c_{kj}) \tag{1}$$

其中，权重计算可表示为：

$$\alpha_{gk}(\dot{x})=\frac{e^{w_{gk}^{\mathrm{T}}\dot{x}_i+b_{gk}}}{\sum_{s=1}^{K}e^{w_{gs}^{\mathrm{T}}\dot{x}+b_{gs}}} \tag{2}$$

$$\alpha_g(\dot{x}_i)=\sigma(w_g^{\mathrm{T}}\dot{x}_i+b_g) \tag{3}$$

其中，σ 是 Sigmoid 函数，$a_{gk}(\dot{x})$ 是计算 $\tilde{x}_i^g$ 与聚类中心 k 的权重值，$\alpha_g(\dot{x}_i)$ 是计算特征所属组的权重值。

整个视频的编码特征可以表示为：

$$y_{jk}=\sum_{i,j}v_{ijk}^g \tag{4}$$

然后，将视频整体特征通过全连接层后输入门控单元。

3.3　门控单元

门控单元 $Z=f(\boldsymbol{Z}_0)$ 将 d_1 维的输入特征向量 $\boldsymbol{Z}_0$ 转换成新的 d_2 维输出向量 $\boldsymbol{Z}$，其数学过程为：

$$\boldsymbol{Z}_1=W_1\boldsymbol{Z}_0+b_1 \tag{5}$$

$$\boldsymbol{Z}_2=\boldsymbol{Z}_1\otimes\sigma(W_2\boldsymbol{Z}_1+b_2) \tag{6}$$

$$\boldsymbol{Z}=\frac{\boldsymbol{Z}_2}{\|\boldsymbol{Z}_2\|_2} \tag{7}$$

其中，$W_1\in\boldsymbol{R}^{d_2\times d_1}$，$W_2\in\boldsymbol{R}^{d_2\times d_2}$，$b_1\in\boldsymbol{R}^{d_2}$，$b_2\in\boldsymbol{R}^{d_2}$ 为可学习参数，σ 是 Sigmoid 激活函数，$\otimes$ 表示逐元素相乘。由式(5)可知，第一层将输入向量 $\boldsymbol{Z}_0$ 映射到一个新的向量空间 $\boldsymbol{Z}_1$。由式(6)可知，第二层是一个内容控制层，通过 $\sigma(W_2\boldsymbol{Z}_1+b_2)$ 学习特征对应的权重，然后与 $\boldsymbol{Z}_1$ 向量逐元素相乘，其中 W_2 和 b_2 是可学习参数。门控单元有两方面作用：1)通过 $\boldsymbol{Z}_1$ 引入非线性；2)赋予 $\boldsymbol{Z}_1$ 向量不同的权重值，类似于注意力机制。由式(7)可知，本文对 $\boldsymbol{Z}_2$ 向量做 L2 正则得到最后的输出向量 $\boldsymbol{Z}$。

3.4　损失函数

在模型训练中，采用双向排序损失函数[19-22]学习文本-视频召回任务。对于 $\forall i\in[1,B]$，视频 Y_i 与其对应的文本 X_i 之间的相似度 $s_{i,i}=s(X_i,Y_i)$ 大于任意其他组合对的相似度 $s_{i,j}$ 和 $s_{j,i}$，其中 $j\neq i$。对于每个 Batch，文本-视频对 $(X_i,Y_i)_{i\in[1,B]}$ 的损失值为：

$$l=\sum_{i=1}^{B}\sum_{j\neq i}[\max(0,m+s_{i,j}-s_{i,i})+\max(0,m+s_{j,i}-s_{i,i})] \tag{8}$$

其中，$s_{i,i}=s(X_i,Y_i)$ 是文本 X_i 和视频 Y_i 的相似度值，m 是间隔值。在本次实验中，$m=0.2$。

3.5　文本-视频的相似度计算

鉴于视频中包含多种模态信息，本文的网络模型分别学习每种模态与文本的相似度，然后将每种模态的相似度值累加后作为文本-视频的相似度值。

首先，计算模态 I_i 和文本 X 的相似度 s_i：

$$s_i(X,I_i)=\langle f_i(h(X)),g_i(h_i(I_i))\rangle \tag{9}$$

其中，$f_i(h(X))$ 表示通过聚类网络 $h(\ )$ 和门控函数 $f_i(\ )$ 得到的向量；$g_i(h_i(I_i))$ 表示视频模态 I_i 经过聚类网络 h_i 和门控函数 g_i 处理后得到的向量；$\langle a,b\rangle$ 表示点乘相加。对于不同的模态 i，该网络会学习到对应的文本向量 $\boldsymbol{f}_i$，从而使每种模态之间具有差异化。

文本-视频的最终相似度值 $s_i(X,I_i)$ 是组合各种模态和文本的相似度得到的：

$$\begin{cases} s(X,Y)=\sum_{i=1}^{N} w_i(X)s_i(X,I_i) \\ w_i(X)=\dfrac{e^{h(X)^T a_i}}{\sum_{j=1}^{N} e^{h(X)^T a_j}} \end{cases} \tag{10}$$

其中，$w_i(X)$是根据文本 X 学习得到的各个模态的权重，$h(X)$是文本特征向量，$a_i(i=1,\cdots,N)$是可学习参数。当某种模态与文本语义强相关时，学到的权重较大，这种方式类似于注意力机制，可以差异化各种模态的重要性，更有利于提高相似的文本-视频的分数值。

当视频中音频或者人脸模态缺失时，将其对应的模态权重置为 0，然后对剩余模态权重做归一化处理：

$$s(X,Y)=\sum_{i\in D}\left[\frac{w_i(X)}{\sum_{j\in D} w_j(X)}\right]s_i(X,I_i) \tag{11}$$

其中，$D\subset\{1,\cdots,N\}$指有效的输入模态类别。

4 实验与结果分析

4.1 实验数据集

MPII 数据集是 LSMDC 比赛的官方数据集，其中包含了从 202 部电影中截取的 118081 个短视频，每个短视频均有 1 个视频描述。MSR 数据集包含 1000 个短视频，每个短视频包含 20 个视频描述。COCO 数据集中的部分图片还有对应的描述，选择该部分图片用于本次实验。本文实验利用视频描述来召回相关视频，评价指标为召准率，即前 K 个召回视频中有正确的视频则认为召回正确。

实验训练的优化函数为 ADAM[23]；在 MPII 数据集上学习率为 0.0001，Batch 大小为 512；在 MSR-VTT 数据集上学习率为 0.0004，Batch 大小为 64。

4.2 实验结果

首先通过对比实验验证人脸模态和增加图片数据集对模型精度的影响，实验结果如表 1 所列，实验数据集为 MPII 数据集。

表 1 模态对比

Table 1 Modal contrast

（单位：%）

Evaluation set / Model	Text-to-Video		
	R@1	R@5	R@10
Our	12.6	27.3	37.2
Our+Face	14.1	29.5	40.3
Our+COCO	13.4	29.2	39.6
Our+Face+COCO	15.2	30.4	40.9

由表 1 可知，人脸模态和在训练数据中增加图片样本可以提升模型精度。

将 COCO 数据集加入 MPII 训练集中，得到的模型精度如表 2 所列。

表 2 MPII 数据集上模型的精度

Table 2 Model accuracy on MPII dataset

（单位：%）

Evaluation set / Model	COCO images			MPII Videos		
	R@1	R@5	R@10	R@1	R@5	R@10
Our+Face	12.6	31.4	45.2	14.1	29.5	40.3
Our+Face+COCO	32.8	66.4	81.3	15.2	30.4	40.9

将 COCO 数据集加入 MSR-VTT 训练集中，得到的模型精度如表 3 所列。

表 3 MSR-VTT 数据集上模型的精度

Table 3 Model accuracy on MSR-VTT dataset

（单位：%）

Evaluation set / Model	COCO images			MPII Videos		
	R@1	R@5	R@10	R@1	R@5	R@10
Our+Face	10.6	27.3	42.7	15.3	39.7	54.2
Our+Face+COCO	23.5	57.4	76.5	19.5	43.8	57.1

表 4 列出了各个模型在 MPII 测试集上的精度对比。

表 4 模型精度对比

Table 4 Model accuracy contrast

（单位：%）

Evaluation set / Model	Text-to-Video		
	R@1	R@5	R@10
Random baseline	0.1	0.5	1.0
C+LSTM+SA+FC7[24]	4.2	13.0	19.5
SNUVL[22]	3.6	14.7	23.9
CT-SAN[2]	5.1	16.3	25.2
Miech et al[25]	7.3	19.2	27.1
CCA(FV HGLMM)[26]	7.5	21.7	31.0
JSFusion[27]	9.5	25.1	38.9
MEE[28]	12.7	25.1	39.6
Our	15.1	28.2	43.4

由表 4 可知，与已有模型相比，本文模型的 $R@1$ 提高了 2.4%，$R@5$ 提高了 3.1%，$R@10$ 提高了 3.8%，在召回任务中表现良好。

4.3 实验分析

通过上述多组实验结果可知，本文提出的模型具有很好的召回效果。该模型支持模态信息缺失，因此可以充分利用图片和视频数据集来同时训练模型，提升模型的精度。同时，本文模型具有很好的扩展性，支持各种模态信息的增减，且所使用的聚类网络也具有很好的将局部信息聚合为全局信息的能力。

结束语 本文提出的模型能很好地将文本-视频映射到一个新的向量空间，使语义相近的文本-视频距离较近，而语义无关的文本-视频距离较远，可以很好地用于文本-视频召回和视频-文本召回任务中。后续将继续探究是否可以增加其他模态信息来提高模型精度，同时也会探索将局部特征信息更好地聚合成全局信息的方法，使全局信息能更好地表征视频信息，从而提高模型精度。

参 考 文 献

[1] TAPASWI M, ZHU Y, STIEFELHAGEN R, et al. Movieqa: Understanding stories in movies through question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[2] YU Y, KO H, CHOI J, et al. End-to-end concept word detection for video captioning, retrieval, and question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[3] PAN Y, MEI T, YAO T, et al. Jointly modeling embeddingand translation to bridge video and language[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[4] PLUMMER B A，BROWN M，LAZEBNIK S. Enhancing video summarization via vision-language embedding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[5] XU R，XIONG C，CHEN W，et al. Jointly modeling deepvideo and compositional text to bridge vision and language in a unifiedframework[C]// Proceeding of the Association for the Advance of Artificial Intelligence. 2015.

[6] YU H，WANG J，HUANG Z，et al. Video paragraphcaptioning using hierarchical recurrent neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[7] TRAN D，BOURDEV L，FERGUS R，et al. Learningspatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015.

[8] CARREIRA J，ZISSERMAN A. Quo vadis，action recognition? a new model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[9] XU J，MEI T，YAO T，et al. Msr-vtt：A large video description-dataset for bridging video and language[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[10] ROHRBACH A，ROHRBACH M，TANDON N，et al. A dataset formovie description[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[11] DENG J，DONG W，SOCHER R，et al. Imagenet：A large-scale hierarchical image database[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009.

[12] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft coco：Common objects in context[C]// Proceedings of the European Conference on Computer Vision. 2014.

[13] KRISHNA R，ZHU Y，GROTH O，et al. Visual genome：Connecting language and vision usingcrowdsourced dense image annotations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[14] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation ofword representations in vector space[C]// Proceedings of the Conference of the Computer and Language. 2013.

[15] ARANDJELOVIC R，GRONAT P，TORII A，et al. NetVLAD：CNN architecture for weakly supervised place recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[16] HE K，ZHANG X，REN S，et al. Deep Residual Learning for ImageRecognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[17] CARREIRA J，ZISSERMAN A. Quo vadis，action recognition? a newmodel and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[18] HERSHEY S，CHAUDHURI S，ELLIS D P W，et al. CNN architectures for large-scale audioclassification[C]// Proceedings of the International Conference on Acoustics，Speech and Signal Processing (ICASSP). 2017.

[19] WANG L，LI Y，LAZEBNIK S. Learning deep structure-preservingimage-text embeddings[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[20] WANG L，LI Y，HUANG J，et al. Learning two-branch neural-networks for image-text matching tasks[C]// Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018.

[21] KARPATHY A，JOULIN A，LI F F. Deep fragment embeddings forbidirectional image sentence mapping[C]// Proceedings of the Conference and Workshop on Neural Information Processing Systems. 2014.

[22] YU Y，KO H，CHOI J，et al. Video captioning and retrieval models with semantic attention[C]// Proceedings of the European Conference on Computer Vision. 2016.

[23] KINGMA D P，BA J. Adam：A method for stochastic optimization[C]// Proceedings of the International Conference on Learning Representations. 2015.

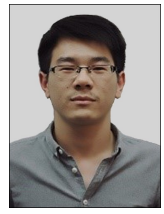
[24] TORABI A，TANDON N，SIGAL L. Learning language-visual embedding for movie understanding with natural-language[C]// Proceedings of the IEEE International Conference on Computer Vision. 2016.

[25] MIECH A，ALAYRAC J B，BOJANOWSKI P，et al. Learning from Video and Text via Large-Scale Discriminative Clustering [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017.

[26] KLEIN B，LEV G，SADEH G，et al. Associating neural word embeddings with deep image representations using fisher vectors [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[27] YU Y，KIM J，KIM G. Joint sequence fusion model for video question-answering and retrieval[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017.

[28] MIECH A，LAPTEV I，SIVIC J. Learning a Text-Video Embedding from Incomplete and Heterogeneous Data[C]// Proceedings of the IEEE Computer Vision and Pattern Recognition. 2019.



ZHANG Heng，born in 1994，master，is a member of China Computer Federation. His main research interests include image processing and deeping learning.

MA Ming-dong，born in 1964，Ph. D，professor，master supervisor. His main research interests include GIS platform software design and development.