基于增强型选择性内核网络的幻灯片切换检测

管民皇 （上海大学通信与信息工程学院,上海 200444）

摘要:主要针对多种类型的演讲视频进行幻灯片切换检测,其主要研究工作包括:提出了一种基于选择性内核模块和全局上下文模块的增强型选择性内核网络,用于进一步提高演讲视频幻灯片切换检测的准确性。选择性内核模块通过分裂、融合、选择三个部分,调整卷积核权重,提取幻灯片语义信息。全局上下文模块则采用压缩扩充模块和非局部模块,调整通道权重、突出有效信息,又捕获长距离依赖、提取幻灯片的全局信息,增强选择性内核网络看清全局的能力。实验结果表明,增强型选择性内核网络用于检测幻灯片切换,其准确性进一步提升,并较好地解决幻灯片之间发生微小变化时的漏检问题。

关键词:演讲视频;幻灯片切换;深度学习;全局上下文模块;选择性内核模块

Abstract:This paper mainly studies the detection of slide transition of speech video based on deep learning.The main research work includes:this paper proposes an Improved Selective Kernel Network based on the selective kernel module and the global context module for slide transition detection of lecture videos.The selective kernel module adjusts the weight of the convolution kernel by splitting,merging,and selecting three parts,and extracts the slide semantic information.The global context module uses compression and expansion modules and non-local modules to adjust channel weights,highlight effective information,capture long-distance dependencies,and extract global information of slides,improving the ability of selective kernel networks to see the whole world.Experimental results show that the Improved Selective Kernel Network can better solve the problem of missed detection when small changes occur between slides.

Keywords:lecture video,slide transition,deep learning,global context module,selective kernel module

现在的演讲视频数目数以万计，很小的错误率乘以大量的演讲视频数目，也会产生影响，提高幻灯片切换检测结果很重要。现在幻灯片切换检测在背景相似,幻灯片切换检测的准确率不高,发生人和幻灯片变化,幻灯片切换检测的准确率不高的问题已经得到解决。本文主要解决幻灯片间发生微小变化,网络模型切换检测准确率不高的问题。图 1 为幻灯片发生微小变化的情况,卷积神经网络错误地判断为幻灯片间没有发生切换。这主要是由于现有的基于深度学习的幻灯片切换检测的，提取图像语义信息的能力有限,提取幻灯片全局信息,理解幻灯片内容的能力有待加强。

图 1 幻灯片发生微小变化

本文提出了选择性内核网络以解决上述问题。选择性内核模块调整不同尺寸的卷积核的特征信息的权重,突出有效信息,抑制无效信息,但只能局限于局部的语义信息。全局上下文模块可以提取全局上下文特征,引入全局信息,看清全局。这两个模块能更好地提取全局语义信息,检测出幻灯片间的微小变化。我们将选择性内核模块和全局上下文模块结合，得到了选择性内核网络,用于演讲视频中的幻灯片切换检测,解决了幻灯片间微小变化难以检测的问题,提高了幻灯片切换检测的准确率。

1 相关技术的发展

深度学习日益发展,在图像处理、视频处理等领域取得了令人瞩目的进步，从卷积神经网络中提取的特征也可以用于幻灯片切换检测。目前将卷积神经网络用于幻灯片切换检测的研究成果很少,所以,这里我们先从现有的卷积神经网络对幻灯片切换检测的效果进行分析，然后再对基于深度学习的幻灯片切换检测的研究现状进行分析。

毫无疑问,3D 卷积神经网络[1]更令人瞩目。3D 卷积神经网络在网络架构、网络模块、相应的损失函数方面都取得了快速发展。这些改进网络架构可以提取到图像的高级特征信息,比传统算法提取到的特征更好,大大提高了幻灯片切换检测的效果。视觉几何组网络(Visual Geometry Group Network,VGGNet)[2]是轻量化网络架构，可以节省大量的训练时间。孪生网络(Siamese Network)[3]将输入信息通过卷积网络映射为特征向量,每次输入需要两个样本作为一个样本对，使用两个向量之间的距离来表示输入之间的差异。文献[4]提出了残差网络(ResNet),提高了特征的重用率，解决了网络模型的梯度消失或梯度爆炸的问题。深度金字塔残差网络[5]使用加法金字塔逐步增加维度,还用了零填充直连的恒等映射,网络更宽、准确率更高。接着,文献[6]提出了增强残差网络(ResNext),改变一个块中相同分支的数量,提高了模型性能。宽度残差网络(WideResNet)[7]改变了卷积神经网络一味增加网络深度的思路，考虑增加网络每一层的深度,有效避免过拟合,提高了网络模型的性能。文献[8]提出了密集网络 (DenseNet),DenseNet 把每个层与它前面的其他层连接起来,特征的传播性好,特征的复用率高,降低了参数量和计算量。

注意力机制的发展也取得了很大的进步。网络架构中加入注意力模块,可以更好地提取幻灯片中的语义信息,理解幻灯片内容,更好地提高幻灯片切换检测的效果。文献[9]提出了残差注意力网络(Residual Attention Network),残差注意力网络使用堆叠注意力结构的方法来改变特征的注意力权重;文献[10]提出双注意力网络(Dual Attention Network),双注意力网络主要有两个注意力模块,自适应地集成局部特征和全局特征。空间通道压缩扩充网络(Spatial-Channel Sequeeze&Excitation,scSE)[11]是对压缩扩充网络(Squeeze-and-Excitation Networks,SENet)[12]的改进,先是压缩空间信息,获得衡量通道重要性的指标;然后压缩通道信息，获得衡量空间信息重要性的指标，最后将这两者相加,对空间信息和通道信息重要性都做了权重处理,突出了有效的特征信息。

2 增强型选择性内核网络检测幻灯片切换

在本节中,我们分别对选择性内核模块,全局上下文模块进行分析，体现这两个模块在提高幻灯片切换检测效果中所起的

重要作用,突出增强型选择性内核网络的优越性。选择性内核模块使用调整不同尺寸卷积核权重的方法提取语义信息，能更好地提取幻灯片的特征信息。理解幻灯片内容,幻灯片切换检测的准确率得以提高。全局上下文模块能够提取全局上下文特征,加强卷积神经网络对全局信息的理解，提高卷积神经网络进行幻灯片切换检测的效果。

2.1 选择性内核模块

图 2 给出的是选择性内核模块[13],选择性内核模块有效地提取了幻灯片的语义信息,该模块主要由三个部分组成,即:

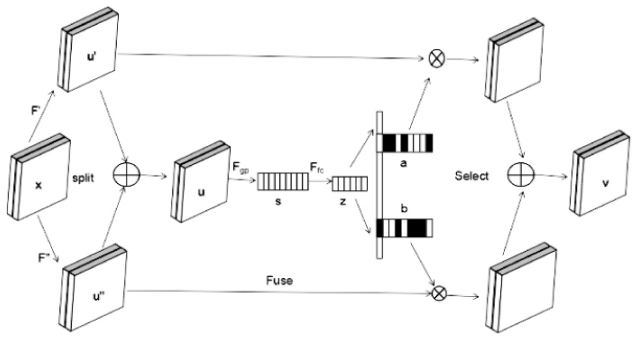

图 2 选择性内核模块

分裂:这部分的主要作用是通过使用不同尺寸的卷积核,将输入的特征分为几个不同的形状大小完全相同的特征，在幻灯片切换检测实验中使用了 2 个卷积核，分别是 1*3*3 和 1*5*5 的卷积。这些卷积都是分组的形式。使用分组卷积可以降低参数数量,使模型更加轻量化。

融合：这部分的主要作用是将不同尺度的特征送入到下一层神经元。主要进行如下操作:

1)各个分支的特征进行元素间的相加:

$$U=U'+U'' \tag{1}$$

2)使用全局平均池化来获得提取特征的全局信息:

$$s_c=F_{gp}(U_c)=\frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W}U_c(i,j) \tag{2}$$

3)通过一个缩减维度的全连接操作,生成一个压缩后的特征向量 $z\in R^{d*1}$。具体为:

$$z=F_{fc}(s)=\delta(\beta(w_s)) \tag{3}$$

其中,δ 指的是 ReLU 激活函数,β 指的是 BN 操作,$w\in R^{d*c}$,文中使用缩减比例系数 r 来控制缩减后的维度 d 对最后网络性能的影响:

$$d=\max(C/r,L) \tag{4}$$

L 指的是网络中 d 能够达到的最小值，在幻灯片切换检测实验中,值为 32。

选择：这部分的主要作用是将融合得到的向量重新分为多个不同的特征向量。幻灯片切换检测实验中,融合得到的向量被分为两个特征向量。接着分别与相应的分裂之后的特征向量进行相应通道的相乘操作，最后再通过这种加权共同构成输入到下一个神经元的特征向量。

$$a_c=\frac{e^{A_c z}}{e^{A_c z}+e^{B_c z}} \tag{5}$$

$$b_c=\frac{e^{B_c z}}{e^{A_c z}+e^{B_c z}} \tag{6}$$

其中,$A,B\in R^{c*d}$,a、b 分别表示 U′、U″的平滑注意力向量。$A_c\in R^{1*d}$ 表示 A 的第 c 行,a_c 表示 a 的第 c 个元素值;B 同理。最终的特征向量 V 表达式为:

$$V_c=a_c\cdot U_c'+b_c\cdot U_c'',a_c+b_c=1 \tag{7}$$

其中,$V=[V_1,V_2,\cdots,V_c],V_c\in R^{H*W}$。

分裂部分形成多尺度特征信息。融合部分将多尺度信息融合,得到对应的选择权重。选择部分根据对应的选择权重对不同尺度特征进行整合。这样做可提取到更有效的语义信息,抑制无用的信息,提高幻灯片切换检测的准确率。

2.2 全局上下文模块

全局上下文模块[14]由压缩扩充模块和非局部模块组成,可以提取幻灯片的全局上下文信息，提高幻灯片切换检测的准确率。全局上下文模块结合了压缩扩充模块和非局部模块的优点,既对不同通道进行权值重标定,调整了通道依赖,又有效地对全局上下文进行了建模，提取到了通道注意力信息和全局上下文信息,不仅可以提取幻灯片的全局信息。还相对轻便,可以集成到卷积神经网络中。如图 3 所示。

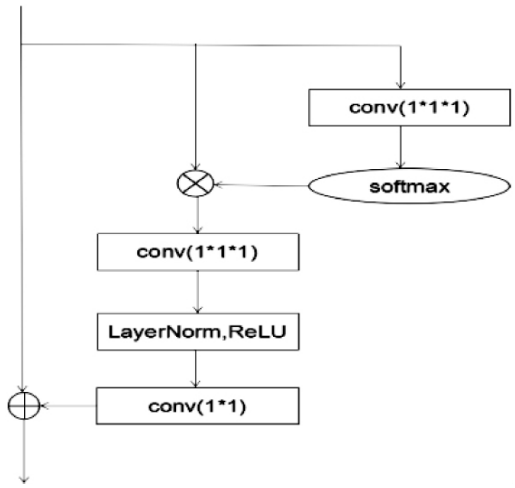

图 3 全局上下文模块

全局上下文模块的表达式为:

$$z_i=x_i+W_{v2}\mathrm{ReLU}(\mathrm{LN}(W_{v1}\sum_{j=1}^{N_p}\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p}e^{W_k x_m}}x_j)) \tag{8}$$

上面的表达式中,x 表示输入,z 表示输出,i、j 表示输入的某个空间位置。

W_k 是 1＊1＊1 的卷积,W_v 是 1＊1＊1 的卷积。在 ReLU 前面增加一个 layer norml-ization 层,降低优化难度,提高了模型的泛化性。

2.3 选择性内核网络结构

选择性内核网络结构如表 1 所示。SK 是选择性内核模块,M 表示选择性内核模块的分裂过程中分支的个数,G 表示分组卷积的分组数目,r 表示选择性内核模块的融合过程中使用的缩

表 1 选择性内核网络结构

层名	相关参数
Conv1	1*3*3,(3,64)
Conv2_x	GC[(64,64)] Sk[M=2,G=8,r=2,(64,256)] Sk[M=2,G=8,r=2,(256,256)] Sk[M=2,G=8,r=2,(256,256)]
Conv3_x	GC[(256,256)] Sk[M=2,G=8,r=2,(256,512)] GC[(512,512)] Sk[M=2,G=8,r=2,(512,512)] GC[(512,512)] Sk[M=2,G=8,r=2,(512,512)]
Conv4_x	GC[(512,512)] Sk[M=2,G=8,r=2,(512,1024)] Sk[M=2,G=8,r=2,(1024,1024)] Sk[M=2,G=8,r=2,(1024,1024)]
	2*7*7，全局平均池化,1024-d 全连接层

减比例系数,GC 是全局上下文模块。M、G、r 这些参数的取值要根据网络模型的大小发生变化。

3 实验结果与分析

3.1 数据集

演讲视频幻灯片切换检测方向无公开的数据库，我们使用自己的演讲视频库进行实验。演讲视频库有六种类型的演讲视频，不仅演讲视频视频类型齐全，还包含了幻灯片内容发生变化,幻灯片内容未发生变化,镜头从幻灯片变化到了人这些变化类型。各类演讲视频的特点为:类型 1,由幻灯片组成;类型 2,由幻灯片和演讲者组成,它们出现在同一屏幕上;类型 3,演讲者遮挡了屏幕;类型 4,演讲者在屏幕附近,没有遮挡屏幕;类型 5,包含了摄像机运动;类型 6,包含了突然的摄像机镜头切换。

3.2 实验设置

把视频逐帧切割,每两帧放入卷积注意力机制双路径网络。最终的幻灯片检测网络分为三类:①幻灯片内容发生了切换;②摄像机镜头从演讲者到幻灯片进行切换；③幻灯片内容没有发生切换。每个视频帧都调整到 112 * 112,被送入网络。

公式(9)~公式(11)用于判断幻灯片切换检测的结果。F-score 是最终的评估指标。

$$precision=\frac{S_c}{S_t} \tag{9}$$

$$recall=\frac{S_c}{S_a} \tag{10}$$

$$F_1=\frac{2*precision*recall}{precision+recall} \tag{11}$$

其中 S_a 是演讲视频库中幻灯片切换的总数,S_c 是正确检测到幻灯片切换的总数，而 S_t 是检测到的幻灯片切换的总数,检测的结果不一定是正确的。

我们将我们提出的方法和其他四种演讲视频镜头切换检测方法进行对比,这四种方法是奇异值分解(SVD)、帧转移参数(FTP)和特征轨迹(SPD)、时空残差网络。其中,奇异值分解(SVD),帧转移参数(FTP)和特征轨迹(SPD)是基于传统方法进行幻灯片切换检测。时空残差网络是基于深度学习的方法进行幻灯片切换检测。

表 2 总结了 SVD、FTP、SPD、时空残差网络、选择性内核网络模型在各类型演讲视频上的平均准确率，平均召回率和平均 F-score 值。Type 5 类演讲视频的 F-score 最低,Type 1 类演讲视频的 F-score 最高。我们的方法的结果比其他三种方法更好,F-score 比 SVD 高 39.7%，比 FTP 高 55.9%，比 SPD 高 21.7%,比时空残差网络高 8.6%。

表 2 四种方法在演讲视频库上的检测结果

	准确率	召回率	F1 分数
SVD	0.726	0.595	0.581
FTP	0.688	0.373	0.419
SPD	0.716	0.840	0.761
STRNet	0.833	0.960	0.892
本文	0.994	0.969	0.978

选择性内核网络是对空间信息的多尺度信息进行权重调整,还提取了全局上下文信息。多尺度信息、全局上下文信息提高了网络的非线性表达能力,使网络有更好的拟合能力,更好地解决了幻灯片切换检测问题。

参考文献

[1]Tran D, Bourdev L, Fergus R,et al. Learnin Spatio temporal Features with 3D Convolutional Networks [C]//IEEE International Conference on Computer Vision, 2015: 4489-4497

[2]Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]//International Conference on Learning Representations , 2015: 1-14

[3]Chopra S, Hadsell R, LeCun Y. Learning a Similarity Metric Discriminatively, with Application to Face Verification[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 539-546

[4]He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition,2016: 770-778

[5]Han D, Kim J, Kim J. Deep Pyramidal Residual Networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5927-5935

[6]Xie S, Girshick R, Doll′ar P, et al. Aggregated residual transformations for deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5987-5995

[7]Zagoruyko S,Komodakis N. Wide residual networks[C]//British Machine Vision Conference,2016: 87.1-87.12

[8]Huang, G.,Liu, Z., Weinberger, K.Q.,Maaten, L.,Densely connected convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition,2261-2269, 2017

[9]Wang F.Residual attention network for image classification[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6450-6458

[10]Zhu Y,Zhao C,Guo H,et al.Attention CoupleNet: fully convolutional attention coupling network for object detection [J]. IEEE Transactions on Image Processing,2019,28(1):113-126

[11]Roy A. G, Navab N, Wachinger C, et al. Concurrent Spatial and Channel ″Squeeze & Excitation″in Fully Convolutional Networks [C]//Lecture Notes in Computer Science, 2018: 421-429

[12]J Hu, L Shen, G Sun. Squeeze-and-excitation networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141

[13]X Li, W Wang, X Hu, et al. Selective Kernel Networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2019: 510-519

[14]Y Cao, J Xu, S Lin, et al. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond [C]//IEEE International Conference on Computer Vision Workshop, 2019: 1971-1980

[收稿日期:2020.7.8]