评述Temporal Relational Reasoning in Videos

舒景东
21821277
浙江大学计算机学院
21821277@zju.edu.cn

摘要（Abstract）

实现物体在时间尺度上的关系的推理是人工智能领域一个重要的问题。活动识别也是计算机视觉领域一个很重要的方向，时序关系推理对于活动识别很重要，因为准确识别一个活动需要模型能够捕获长时间尺度和短时间尺度的物体的关系，而目前存在的其他活动识别的方法基本上基于物体的外观，很难捕获长期和短期的物体时序关系。文章作者提出Temporal Relation Networks（TRN）以及多尺度的思路以用于解决这个问题，同时通过采样的方式减少计算量。本文首先介绍作者提出的Temporal Relation Networks（TRN）以及多尺度的思路，提出对于模型的几点改进，然后分析了实验结果，结果说明了本文提出的改进的有效性。本文实验代码地址：https://github.com/djshu/TRN-pytorch

1. Temporal Relation Networks

Temporal Relation Networks[1]可以当做一个单独的模型用在很多卷积神经网络中，可复用性和扩展性较好，下面首先介绍Temporal Relations，再介绍Multi-Scale Temporal Relations,最后介绍Temporal Relation Networks的效率

1.1. Temporal Relations

受到视觉问答系统中关系推理模块的启发[2],作者提出如下的捕获成对关系的函数：

$$T_2(V)=h_\phi\left(\sum_{i<j} g_\theta\left(f_i,f_j\right)\right) \tag{1}$$

输入是挑选出的有序的n帧视频，经过卷积神经网络的处理得到V，即$V=\{f_1,f_2,...,f_n\}$，其中f_i代表视频里的第i帧经过卷积神经网络得到的特征。对于h_ϕ和g_θ，作者采用MLP，ϕ和θ分别为其参数。为了减少计算量，作者不是对于输入中所有两两组合的情况都使用g_θ计算其relation，而是从输入均匀采样出一部分的两两组合进行计算。

作者也不局限于只提取两帧之间的关系，可以提取更多帧之间的关系，比如提取三帧之间的关系：

$$T_3(V)=h'_\phi\left(\sum_{|i<j<k} g'_\theta\left(f_i,f_j,f_k\right)\right) \tag{2}$$

同样地，三帧都是从输入均匀采样得到的有序的。

1.2. Multi-Scale Temporal Relations

为了捕获不同时间尺度下的时序关系，作者采用如下方式：

$$MT_N(V)=T_2(V)+T_3(V)\ldots+T_N(V) \tag{3}$$

每个T_d用于捕获d帧视频间的时序关系，且对应其$h_\phi^{(d)}$和$g_\theta^{(d)}$。Temporal Relations和Multi-Scale Temporal Relations都可以进行端到端的训练。

1.3. 提高计算效率的方式

对于前面提到的T_d，如果d小于输入的视频的总帧数，则d帧视频有很多种可能的选取方式，特别是当d相比视频总帧数小得多的时候，可能的选取方式更多，如果都进行计算，则计算量太大，所以作者提出首先从输入的视频里采样N帧，得到V_N^*，$V_N^*\subset V$，计算$T_N(V)$。然后对于每个小于N的d，从V_N^*中采样k次，每次采样d帧视频，得到V_{kd}^*用于计算$T_d(V)$，其中$V_{kd}^*\subset V_N^*$。对于每个视频，由于用于提取视频中每帧图像特征的CNN只需要在N帧上进行计算，得到$V_N^*=\{f_1,f_2,...,f_N\}$，后续采样是直接在提取出的特征$\{f_1,f_2,...,f_N\}$中采样，这样就省去了很大一部分CNN提取特征的计算量。

模型测试阶段，等间隔取视频中的帧，输入到CNN提取特征，然后送入队列，然后从队列中采样，计算多尺度的时序关系。得益于这种高效率的设计，作者提出的Temporal Relation Networks可以运行在普通台式机上对电脑摄像头的视频进行实时处理。

1.4. 模型细节设计

作者考虑到模型准确率和效率之间不可能同时满足，所以为了追求准确率和效率的平衡，实验中用于提

取每帧图像特征的CNN都采用在ImageNet上预训练过的带有Batch Normalization的Inception网络(BNInception)[3],且按照论文[4]，除了第一个Batch Normalization层，其他都冻结参数，训练中不更新，且在全局池化层后采用Dropout策略。用于提取时序关系的每帧图像的特征采用全局平均池化层后的值。对于之前提到的每个d的采用次数k，实验中取 3。$g_\theta^{(d)}$ 为双层MLP，每层有 256 个神经元，$h_\phi^{(d)}$ 为单层MLP，神经元数目为数据集的类别数目。对于多尺度TRN，作者考虑从 2 帧到 8 帧的数据，即公式(3)提到的N取 8。

2. 改进模型

由于CNN提取的图像的特征对于时序关系的推断十分重要，ResNet[5]是图像分类领域十分成功的网络框架，也成功地被用作目标检测，语义分割等多种计算机视觉任务的Backbone，本文将采用ResNet101 提取图像特征进行实验并将结果和作者用BNInception的结果进行对比。

作者论文中实现的Baseline之一的VideoLSTM[6]是将每帧视频图像用CNN提取特征然后送入LSTM进行处理，将LSTM的所有时间点的输出求平均得到此视频的类别信息。由于是处理输入视频的所有帧，所以计算效率比较低，且帧数太多，也容易导致视频里无效信息淹没有效信息。由于自然语言领域里LSTM十分适合处理时序信息，所以结合作者TRN的多尺度提取时序信息的思想，本文提出用LSTM替换作者TRN中的 g_θ 进行实验，具体而言，LSTM不采用Dropout机制，hidden state size为 256。

分析作者开源的代码后发现，CNN得到的特征不是直接输入到TRN进行计算，而是经过了一层MLP，且对于多尺度TRN，除了作者论文里的实现以外，还有两个变种。本文进行实验的模型概括如下，其中S1 和S2 为两个变种:

S0: 作者论文里多尺度TRN的实现。

S1: $g_\theta^{(d)}$ 为两层MLP，其中第二层神经元数目为数据集类别数目，不采用 $h_\phi^{(d)}$ 。

S2:对于每d帧视频的k次采样的结果先经过 $h_\phi^{(d)}$ 进行计算再求和，即：

$$T_d(V)=\sum_{i_1<i_2<...<i_d} h_\phi^{(d)}(g_\theta^{(d)}(f_{i_1},f_{i_2},...,f_{i_d})) \quad (4)$$

其中 $g_\theta^{(d)}$ 为双层MLP， $h_\phi^{(d)}$ 为单层MLP, 输出维度为类别数目。

S3: $g_\theta^{(d)}$ 为单向双层LSTM，LSTM最后一层的最后一个time step的hidden state进行求和，然后输入到 $h_\phi^{(d)}$ ，$h_\phi^{(d)}$ 为单层MLP,输出维度为类别数目。 。

$$T_d(V)=h_\phi^{(d)}\left(\sum_{i_1<i_2<...<i_d} g_\theta^{(d)}(f_{i_1},f_{i_2},...,f_{i_d})\right) \quad (5)$$

S4: $g_\theta^{(d)}$ 为单向双层LSTM，LSTM最后一层的最后一个time step的hidden state进行求和，然后输入到 $h_\phi^{(d)}$ ，$h_\phi^{(d)}$ 为单层MLP,输出维度为类别数目。

$$T_d(V)=\sum_{i_1<i_2<...<i_d} h_\phi^{(d)}(g_\theta^{(d)}(f_{i_1},f_{i_2},...,f_{i_d}))$$

$$T_d(V)=\sum_{i_1<i_2<...<i_d} h_\phi^{(d)}(g_\theta^{(d)}(f_{i_1},f_{i_2},...,f_{i_d})) \quad (6)$$

S5: $g_\theta^{(d)}$ 单向单层LSTM，其他和S4 相同。

S6:TSN

S7:提取图像特征的CNN采用ResNet101，其他方面和S1 一致，S0-S6 中用于提取图像特征的CNN都采用BNInception。

3. 实验

3.1. 数据集

作者论文的实验在活动识别这个任务上进行，采用了4 个数据集，分别为Something-V1[7] and Something-V2 [8], Jester dataset [9], and Charades dataset [10]。这些数据集通过众包建立，即视频中的人物按照指令做相应的动作，视频具有明显的起止时刻，所以相比UCF101 和Kinects这些来自YouTube的视频，为了正确判断视频中活动的类型，更加需要时序推理。

数据集	类别数	视频数	类型
Something-V1	174	108499	人和物体交互
Something-V2	174	220847	人和物体交互
Jester	27	148092	手势
Charades	157	9848	室内日常活动

3.2. 实验结果

3.2.1 Something-Something数据集上的结果

Something-Something是用于识别人和物体交互的视频数据集，包括 174 个类别，且有一些含义上比较接近的类别，比如“把东西撕成两半”和“把东西撕开一点点”，即为了区分这些比较接近的类别，需要模型很好地提取视频中的时序信息。Something-Something数据集

上的结果如表 1所示，Something-V1 的结果为Top1 Accuracy，Something-V2 上的结果为Top1 和Top5 Accuracy，Baseline指的是从每个视频中随机选择一帧进行计算，2-stream TRN指对每个视频的两个stream得到的分类结果进行平均，作为最终的预测结果。可以看到，2-stream TRN相比单stream的Multiscale方法效果更好。

为了验证时序的重要性，作者还通过实验比较了TRN和TSN[4]。TSN的方法是对于视频中的每帧图像提取特征得到相应的每个类别的概率，或者说class score，然后对不同的帧的class score求平均得到最终的预测结果，这种方法只能突出提取不同帧中同时出现的信息，而无法提取时序信息。作者实验时，保持其他实验条件不变，用不同的帧数进行实验，得到的Something-V1 的验证集上的结果如表 2所示。可以得知，不同帧数下，TRN效果都比TSN好，且随着帧数的上升，TRN的效果提升显著，而TSN提升很小，说明帧数上升提供的更多的时序信息能够被TRN捕获从而提升模型效果。

表 1 Something数据集结果

	Something-V1		Something-V2	
	Val	Test	Val	Test
Baseline	11.41	-	-	-
MultiScale TRN	34.44	33.60	48.80/77.64	50.85/79.33
2-Stream TRN	42.01	40.71	55.52/83.06	56.24/83.15

表 2 TRN和TSN结果对比

	TRN	TSN
2-fr.	22.23	16.72
3-fr.	26.22	17.30
5-fr.	30.39	18.11
7-fr.	31.01	18.48

3.2.2 Jester和Charades数据集上的结果

Jester验证集上结果和测试集上结果分别如表 3和表 4所示，验证集上结果显示，随着帧数的增加，TRN的效果依然不断变好，且多尺度效果最好，测试集上效果显示，Jester数据集上多尺度TRN达到了state-of-the-art。

Charades数据集上结果如表 5所示，同样地，此数据集上也是多尺度TRN效果最好，击败了C3D[10]以及Asynchronous Temporal Field (TempField) [11]等方法。

表 3 Jester验证集上结果

	Val
Baseline	63.60
2-frame TRN	75.65
3-frame TRN	81.45
4-frame TRN	89.38
5-frame TRN	91.40
MultiScale TRN	95.31

表 4 Jester测试集上结果

	Test
20BN Jester System	82.34
VideoLSTM	85.86
Guillaume Berger	93.87
Ford's Gesture System	94.11
Besnet	94.23
MultiScale TRN	94.78

表 5 Charades数据集上结果

Approach	Random	C3D	AlexNet	IDT
mAP	5.9	10.9	11.3	17.2
Approach	2-Stream	TempField	Ours	
mAP	14.3	22.4	25.2	

3.3. 探究TRN中的Visual Common Sense Knowledge

为了体现TRN的可解释性，作者分别从以下四个方面分析TRN的Visual Common Sense Knowledge：TRN得到的用于识别活动的代表性的帧；视频的时序对齐；时序对于活动识别的重要性；早期活动识别。

3.3.1 TRN得到的用于识别活动的代表性的帧

作者首先从视频中等距离取不同帧，用CNN计算特征，然后随机组合成不同帧数，再通过TRN计算响应，得到预测活动类别最高响应的帧，在Something数据集的验证集上得到的某些帧如图 1所示。可以发现，TRN可以提取用于识别活动的关键帧，且对于一些简单活动，2 帧甚至 1 帧就能正确识别出活动，但是一些需要更多时序信息的活动，就需要更多帧才能正确识别。

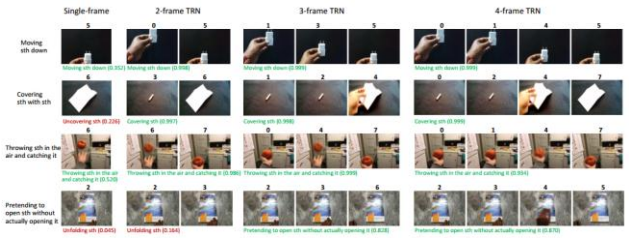

图 1 关键帧示例

3.3.2 视频的时序对齐

由于上一节分析发现对于相同类型活动的不同视频，TRN提取出的关键帧具有一致性，所以TRN可以用来实现视频的时序对齐。具体而言，给定包含相同动作的几个视频，利用TRN提取视频里最能反映动作类型的帧作为时间轴上的标记，然后将相邻两个标记之间的视频内容的帧数调节成一致，视频时序对齐后的例子如图 2所示。而且视频时序对齐这个任务，也是TRN独有的优点，之前的其他方法比如 3D卷积网络，2-stream网络都无法做到。

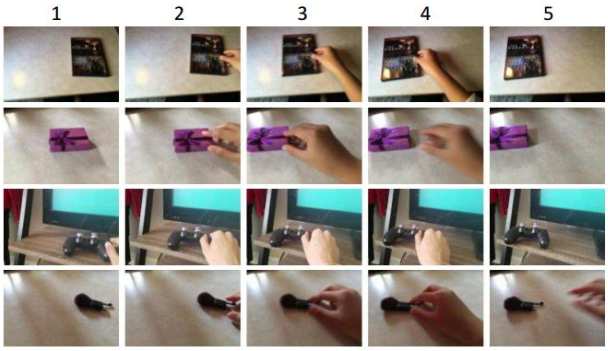

图 2 视频时序对齐示例——将物体从右边推到左边

3.3.3 时序对于活动识别的重要性

为了探究时序对于活动识别的重要性，作者分别在Something-Something和UCF101 数据集[12]上进行实验，分别用有序和随机打乱帧顺序的视频训练TRN，实验结果如图 3 图 4所示。可以发现对于Something-Something数据集而言，乱序后准确率下降很多，而UCF101 数据集乱序后准确下降小，说明识别Something-Something数据集中活动类型相比UCF101 更需要准确的时序信息。

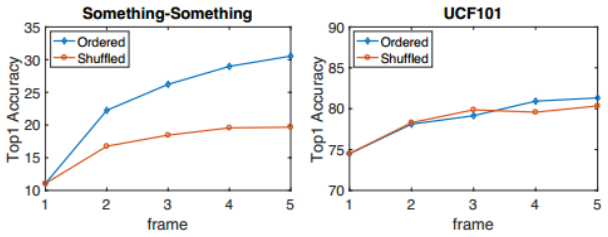

图 3 有序和打乱顺序帧顺序后TRN的Top1 Accuracy

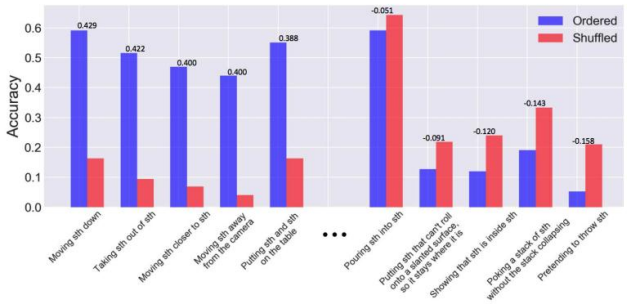

图 4 Something-Something数据集里有序相比乱序准确率增益最大和最小的五类数据

3.3.4 早期活动识别

早期活动识别指在活动发生前或者已知一部分行为时识别出此活动，是活动识别中很有挑战性的任务。为了检测TRN的早期活动识别效果，作者分别将数据的验证集里每个视频的前 25%和 50%的内容输入给TRN进行预测，结果如表 6所示。表中的baseline由视频中随机选择的单帧视频训练而成，可以得知在Something和Jester数据集上多尺度TRN都具有一定的早期活动识别能力。

表 6 多尺度TRN早期活动识别结果

	Something		Jester	
Frames	baseline	TRN	baseline	TRN
first 25%	9.08	11.14	27.25	34.23
first 50%	10.10	19.10	41.43	78.42
full	11.41	33.01	63.60	93.70

3.4. 改进模型的对比实验

实验在Something-Something V1 上进行，视频采用RGB数据，S0-S6 的batch size取 32，由于S7 显存需求量过大，batch size取 16。实验结果如下：

表 7改进模型的对比实验结果

	S0	S1	S2	S3	S4	S5	S6	S7
val top1 acc	29.33	32.54	30.5	32	31.67	30.54	17.95	32.71
test top1 acc/top5 acc	30.96/ 59.96	33.96/ 62.98	31.67/ 61.19	33.48/ 62.51	32.88/ 62.07	31.96/ 60.80	17.84/ 44.58	33.84/ 63.34

分析实验结果可知，两个模型的变种均比原论文中提出的多尺度TRN效果要好，且变种S1 和S7 的区别在于S1 采用BNInception提取图像特征，S7 采用ResNet101 提取图像特征，val top1 准确率和test top5 准确率S7 比S1 高，但是test top1 准确率S1 稍高，说明采用ResNet101 提取图像特征确实能够给模型带来一定提升，尽管提升不大，但是采用ResNet101 的S7 训练和推断时间至少是S1 的两倍，所以作者说权衡准确率和计算效率后采用BNInception提取图像特征还是有道理的。除S6 这个TSN之外的模型都基于作者的多尺度的思想，可以看到准确率远远好于S6，证明作者提出的多尺度的思想的有效性。用LSTM的S3，S4，S5 都有不错的效果，其中S3 十分接近S1，说明本文提出采用LSTM结合作者的多尺度思想也能捕获有用的时序信息。

4. 结论

本文介绍了《Temporal Relational Reasoning in Videos》这篇论文的方法，通过论文在多个数据集上的实验结果分析了其有效性，实用性和可解释性；另外，针对这篇论文提出了几点改进，并通过实验验证了改进的有效性。

References

[1] Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018). Temporal Relational Reasoning in Videos. European Conference on Computer Vision,, 831-846.

[2] Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia,P., Lillicrap, T.: A simple neural network module for relational reasoning. arXiv preprint arXiv:1706.01427 (2017)

[3] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. (2015) 448–456

[4] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: Proc.ECCV. (2016)

[5] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition.(2016) 770–778

[6] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 2625–2634

[7] Goyal, R., Kahou, S., Michalski, V., Materzy ńska, J., Westphal, S., Kim, H.,Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. Proc.ICCV (2017)

[8] Mahdisoltani, F., Berger, G., Gharbieh, W., Fleet, D., Memisevic, R.: Fine-grained video classification and captioning. arXiv preprint arXiv:1804.09235 (2018)

[9] Goyal, R., Kahou, S., Michalski, V., Materzy ńska, J., Westphal, S., Kim, H.,Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. Proc.ICCV (2017)

[10] Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In:European Conference on Computer Vision, Springer (2016) 510–526

[11] Sigurdsson, G.A., Divvala, S., Farhadi, A., Gupta, A.: Asynchronous temporal fields for action recognition. (2017)

[12] Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. Proc. CVPR (2012)