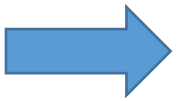

Apply Eye Makeup
Apply Lipstick
Blow Dry Hair
Brushing Teeth
Cutting In Kitchen
Hammering
Hula Hoop
Juggling Balls
Jump Rope
Knitting
Mixing Batter
Mopping Floor
Nun Chucks
Pizza Tossing
Shaving Beard
Skate Boarding
Soccer Juggling
Typing
Writing On Board
Yo Yo
Baby Crawling
Blowing Candles
Body Weight Squats
Handstand Pushups
Handstand Walking
Jumping Jack
Lunges
Pull ups
Push ups
Rock Climbing Indoor
Rope Climbing
Swing
Tai Chi
Trampoline Jumping
Walking with a Dog
Wall Pushups
Band Marching
Haircut
Head Massage
Military Parade
Salsa Spin
Drumming
Playing Cello
Playing Daf
Playing Dhol

视频分类技术整理

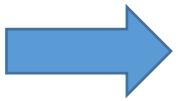

Playing Flute
Playing Guitar
Playing Piano
Playing Sitar
Playing Tabla
Playing Violin
Archery
Balance Beam
Baseball Pitch
Basketball
Basketball Dunk
Bench Press
Biking
Billiard
Bowling
Boxing-Punching Bag
Boxing-Speed Bag
Breaststroke
Clean and Jerk
Cliff Diving
Cricket Bowling
Cricket Shot
Diving
Fencing
Field Hockey Penalty
Floor Gymnastics
Frisbee Catch
Front Crawl
Golf Swing
Hammer Throw
High Jump
Horse Race
Horse Riding
Ice Dancing
Javelin Throw
Kayaking
Long Jump
Parallel Bars
Pole Vault
Pommel Horse
Punch
Rafting
Rowing
Shotput
Skiing

图像分类

H x W x C



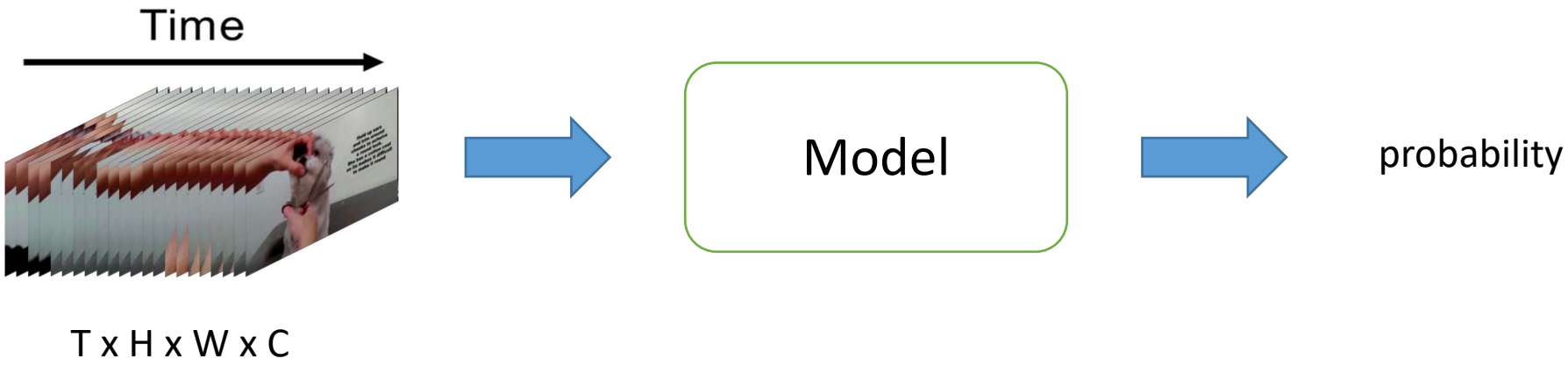
Model



probability

视频分类：2D -> 3D

视频一般可视为由一系列图像帧组成的图像序列



T x H x W x C

- 需要考虑到帧之间的变化信息
- 相邻帧之间存在大量冗余

解决方案

1. 最直接的解决方案

直接使用2D图像的方法，针对每个视频帧，利用神经网络提取其每个视频帧图像的特征，将这个视频所有帧的特征向量取平均得到整个视频的特征向量，然后进行分类识别，或是直接每个帧得到一个预测结果，最后在所有帧结果中取共识

- 损失帧之间的相互关系
- 计算量小，实现简单
- 在视频非常短时效果好

解决方案

2. VLAD

对一个视频的各个帧特征进行聚类得到多个聚类中心，将所有的特征分配到指定的聚类中心中，对于每个聚类区域中的特征向量分别计算，最终concat或加权求和所有的聚类区域的特征向量作为整个视频的特征向量

- VLAD（2010）：

Formally, given N D-dimensional local image descriptors $\{\mathbf{x}_i\}$ as input, and K cluster centres ("visual words") $\{\mathbf{c}_k\}$ as VLAD parameters, the output VLAD image representation V is $K \times D$-dimensional. For convenience we will write V as a $K \times D$ matrix, but this matrix is converted into a vector and, after normalization, used as the image representation. The (j, k) element of V is computed as follows:

$$V(j,k) = \sum_{i=1}^{N} a_k(\mathbf{x}_i)\left(x_i(j) - c_k(j)\right), \qquad (1)$$

N x D 向量 -> K x D 向量

where $x_i(j)$ and $c_k(j)$ are the j-th dimensions of the i-th descriptor and k-th cluster centre, respectively. $a_k(\mathbf{x}_i)$ denotes the membership of the descriptor $\mathbf{x}_i$ to k-th visual word, *i.e.* it is 1 if cluster $\mathbf{c}_k$ is the closest cluster to descriptor $\mathbf{x}_i$ and 0 otherwise. Intuitively, each D-dimensional

解决方案

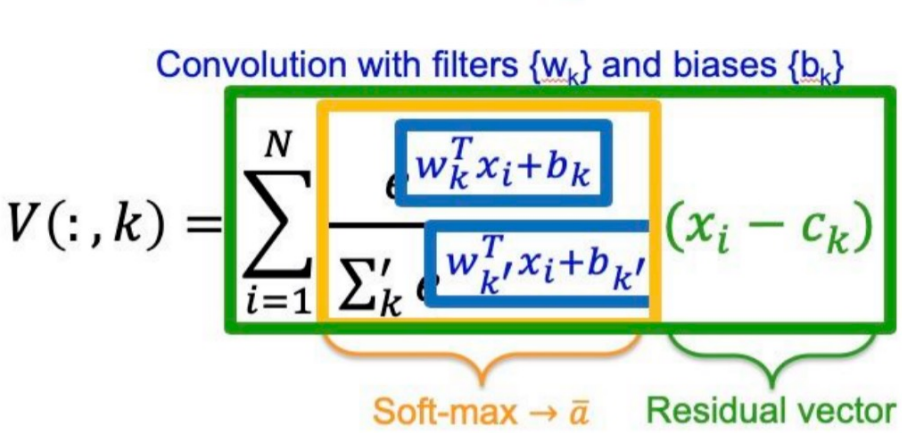
NetVLAD：

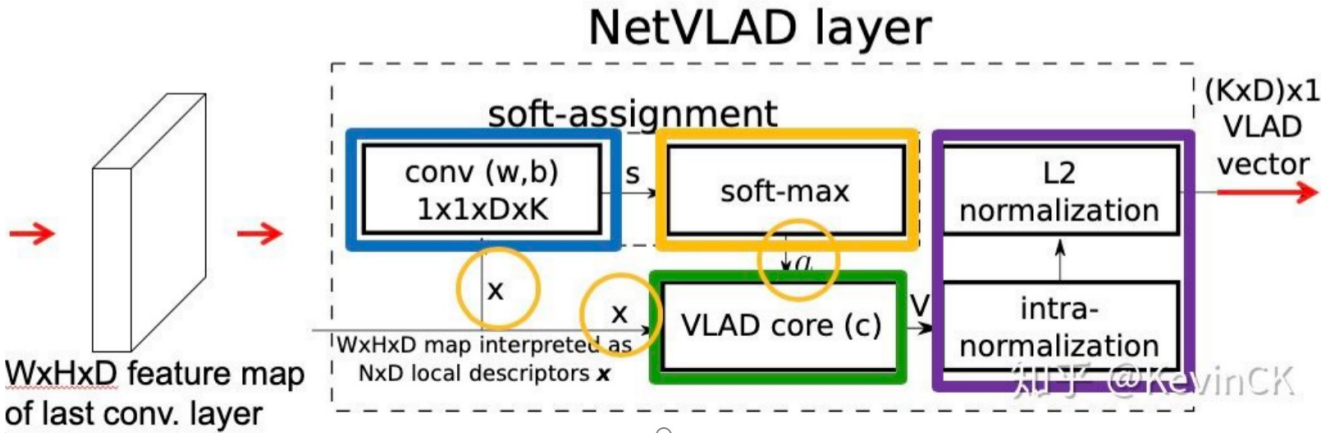
不可导

$$V(j,k) = \sum_{i=1}^{N} a_k(\mathbf{x}_i)\left(x_i(j) - c_k(j)\right),$$

$$\bar{a}(x_i) = \frac{e^{-\alpha||x_i - c_k||^2}}{\sum_{k'} e^{-\alpha||x_i - c_{k'}||^2}}$$

- 相比于Average Pooling，NetVLAD可以通过聚类中心将视频序列特征转化为多个视频镜头特征，然后通过可以学习的权重对多个视频镜头加权求和获得全局特征向量。
- 每一帧的特征向量仍是独立计算的，无法考虑帧之间的时序关系

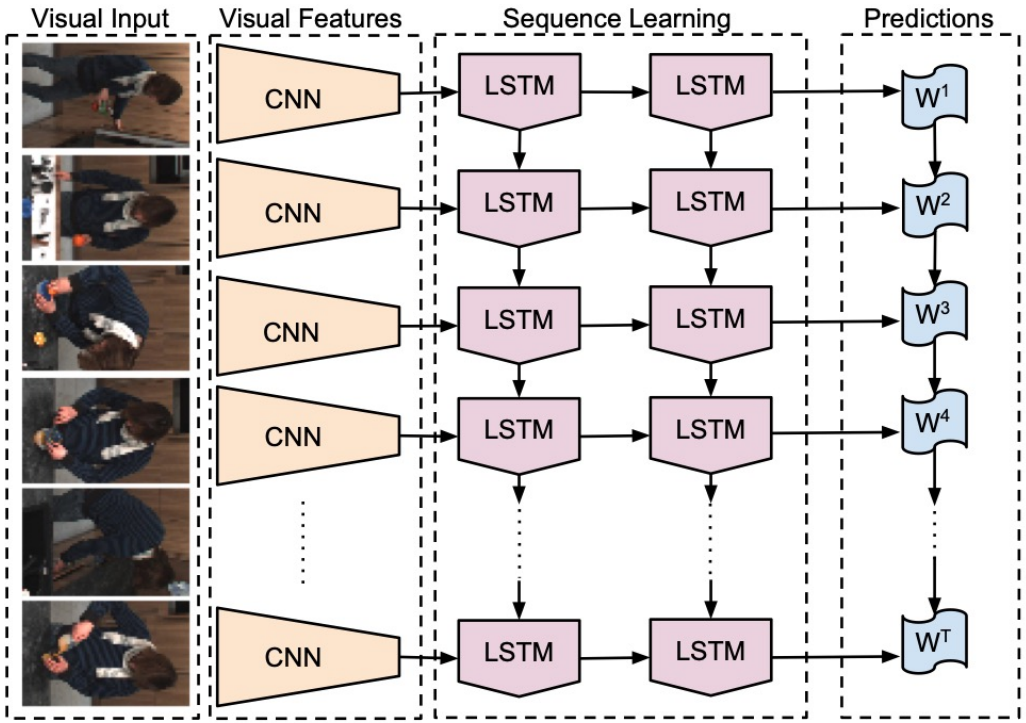




解决方案

3. RNN融合

将由CNN提取到的帧特征作为时序特征序列，输入到RNN网络中。

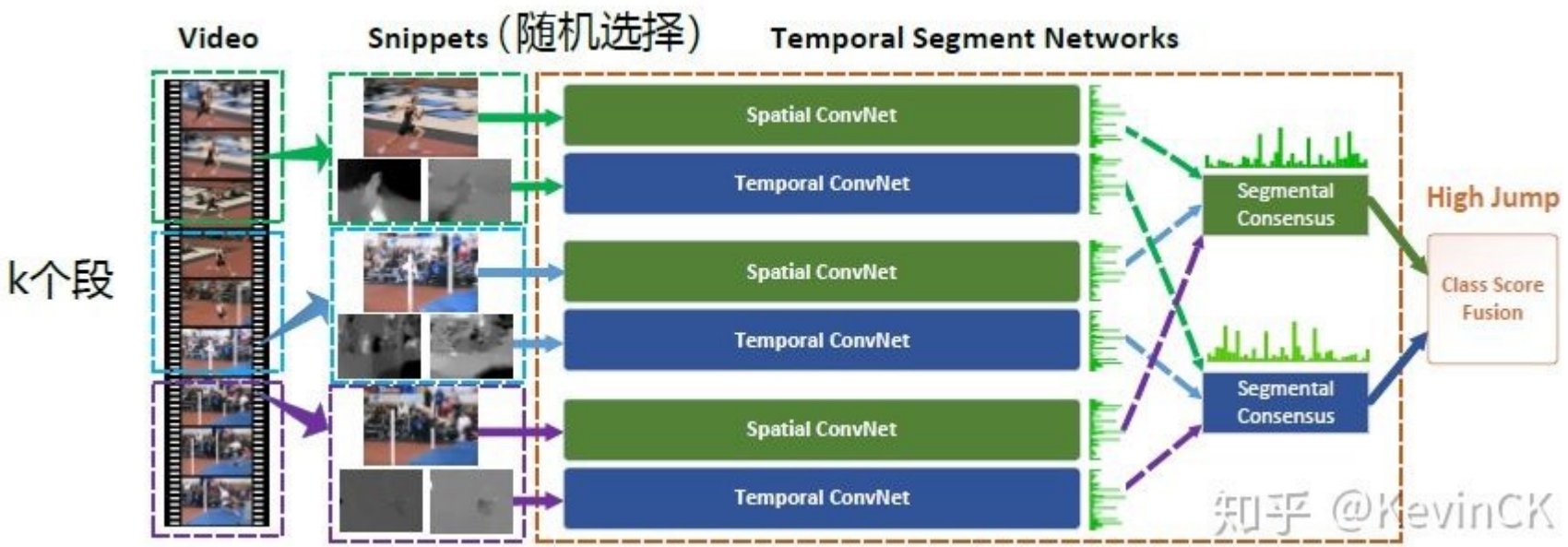


如：Long-term Recurrent Convolutional Networks for Visual Recognition and Description

解决方案

4. 双流法

两个网络分支，一个为图像分支提取视频帧的特征向量，另一个为光流分支，利用多帧之间的光流图提取光流特征，利用图像分支与光流分支特征向量的融合进行分类预测。



如：Temporal Segment Network

解决方案

5. 3D卷积

直接将卷积核扩展为三维，将输入序列看作一个整体

- 实验效果优于基于池化的二维网络
- 模型复杂度高，需要大量数据，训练困难

- 由于3D卷积网络训练困难，衍生出一系列相关变体，如P3D，FSTCN等基于低秩近似的方法

如：C3D（3D版VGG），Res3D（3D版ResNet），Slow-Fast

解决方案

5. 3D Attention

将视觉Transformer中的注意力扩充至3D

- 模型开销更高，ViT本身就属于大模型，扩充至3维后复杂度提高一个数量级

解决方案

TimeSformer：基于ViT的3D Attention扩展

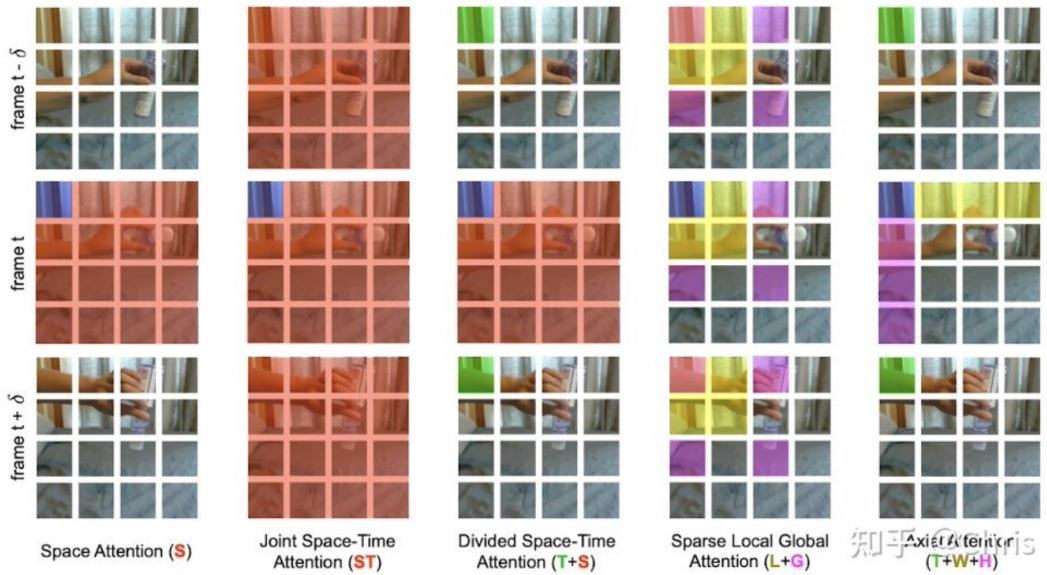
1. 空间注意力机制（S）：只取同一帧内的图像块进行自注意力机制

时空共同注意力机制（ST）：取所有帧中的所有图像块进行注意力机制

2. 分开的时空注意力机制（T+S）：先对同一帧中的所有图像块进行自注意力机制，然后对不同帧中对应位置的图像块进行注意力机制

3. 稀疏局部全局注意力机制（L+G）：先利用所有帧中，相邻的 H/2 和 W/2 的图像块计算局部的注意力，然后在空间上，使用2个图像块的步长，在整个序列中计算自注意力机制，这个可以看做全局的时空注意力更快的近似

4. 轴向的注意力机制（T+W+H）：先在时间维度上进行自注意力机制，然后在纵坐标相同的图像块上进行自注意力机制，最后在横坐标相同的图像块上进行自注意力机制



解决方案

Video Swin Transformer：Swin Transformer的3D拓展