3D Convolutional Neural Networks for Human Action Recognition

Shrikunj Sarda - 2016B1A80812P
Mayank Singh - 2016B1A30859P

Shuiwang Ji, Wei Xu, Ming Yang

Outline

Introduction

- **Applied the 3D convolution operation to extract spatial and temporal features from video data for action recognition.**

3D CNN Architecture (TRECVID)

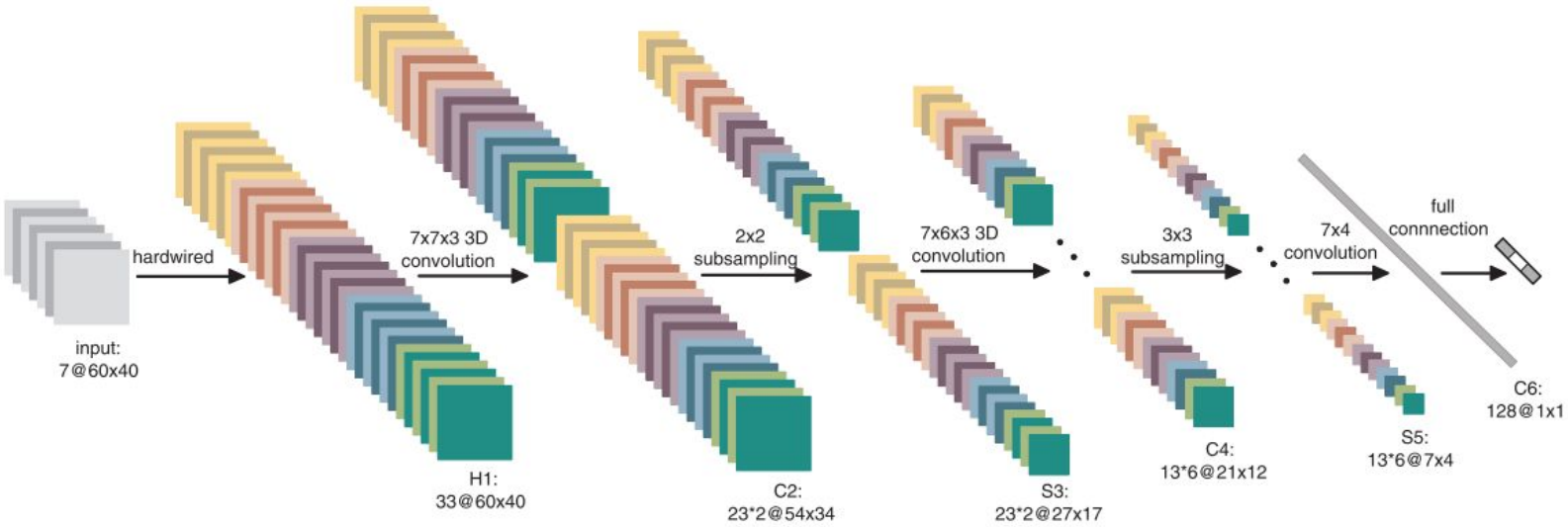


Fig. 3. A 3D CNN architecture for human action recognition. This architecture consists of one hardwired layer, three convolution layers, two subsampling layers, and one full connection layer. Detailed descriptions are given in the text.

3D CNN Architecture(KTH Dataset)

1) 9 frames from each video are extracted
2) 5 channels from these 9 frames are applied
3) Channels are:
    i) grayscale - 9
    ii) gradient x - 9
    iii) gradient y - 9
    iv) optflow x - 8
    v) optflow y - 8

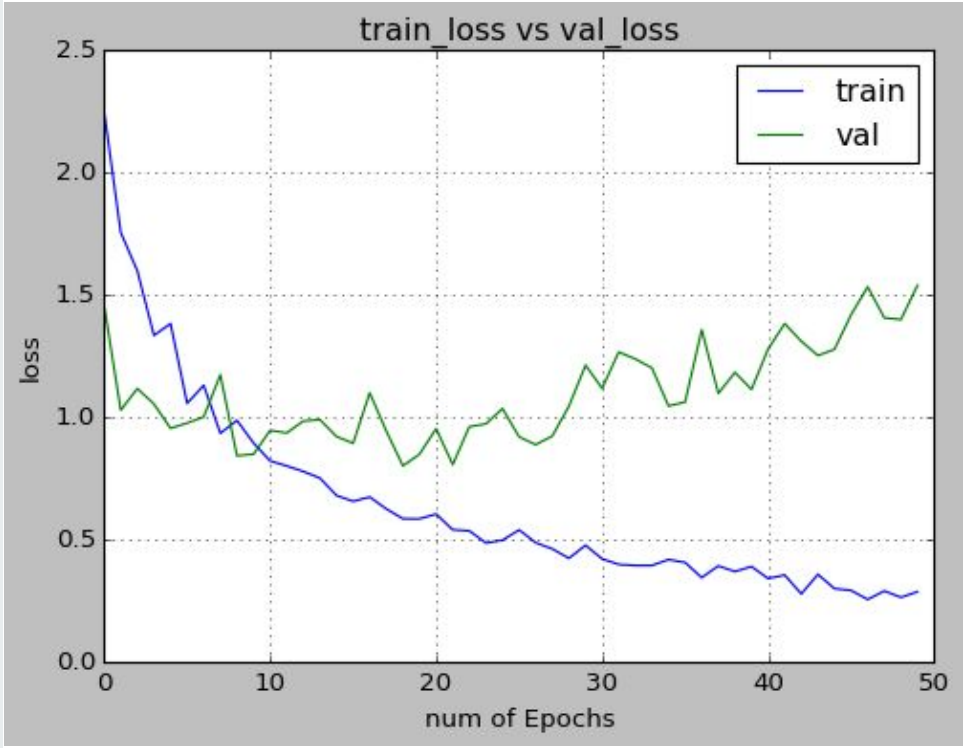
3D CNN Architecture(KTH Dataset)

4) total of 43 frames for input in 3D CNN model
5) 3 Conv layers and 2 maxpool layers are applied
6) first layer is conv with kernel size 9X7X3 followed by 3X3X1 maxpool
7) second layer is conv with kernel size 7X7X3 followed by 3X3X1 maxpool
8) third is conv with kernel size 6X4X1 followed by a dense layer
9) relu is used as activation function for all layers except final layer in which softmax is used
10) data is classified in 6 classes given in KTH dataset

KTH Dataset

- The KTH dataset consists of videos of humans performing 6 types of action: boxing, handclapping, handwaving, jogging, running, and walking.
- There are 25 subjects performing these actions in 4 scenarios: outdoor, outdoor with scale variation, outdoor with different clothes, and indoor.
- The total number of videos is therefore 25x4x6 = 600 (~599 as 1 missing). The videos' frame rate are 25fps and their resolution is 160x120.

Training Loss vs Validation Loss



Training Accuracy vs Validation Accuracy