*2019 Anomaly Detection in Video Sequence with Appearance-Motion Correspondence*

Abstract:

They propose a deep convolutional neural network (CNN) that address the diversity of possible events by learning a correspondence between common object appearances and their associated motions.

Introduction:

They design a CNN that combines a Conv-AE and a U-Net/

They integrate an Inception module after the input layer to reduce the effect of network’s depth since the depth is considered as a hyper parameter that requires a careful selection.

They propose a patch-based scheme estimating frame-level normality score that reduces the effect of noise.

Proposed Method:

The model includes two processing streams: 1. Via Conv-AE to learn common appearance spatial structures in normal events. 2. Determine an association between each input pattern and motion.

Skip connections are not employed in the appearance stream because it could let the input information go through these connections.

The size of input layer: 128 x 192 x 3.

Inception Module:

The inception module was originally proposed to let a CNN decide its filter size automatically.

They apply an inception module to let the model select its appropriate convolutional operations.

They remove the max-pooling and include 4 streams of convolutions of filter sizes 1 x 1, 3 x 3, 5 x 5 and 7 x 7. Each convolutional layer of filter larger than 1 x 1 is factorized into a sequence of layers with smaller receptive fields in order to reduce the computational cost.

Appearance Convolutional Autoencoder:

Encoder and decoder without skip connection.

Encoder: convolution, batch-normalization (BatchNorm) and leaky-RELU activation. The first block does not contain BatchNorm layer. Instead of using pooling layer to reduce the resolution of feature maps, they apply stride convolution.

Decoder: a dropout layer (with 0.3) is attached before the RELU activation as a regularization that reduces the risk of overfitting.

Intensity loss ( distance):

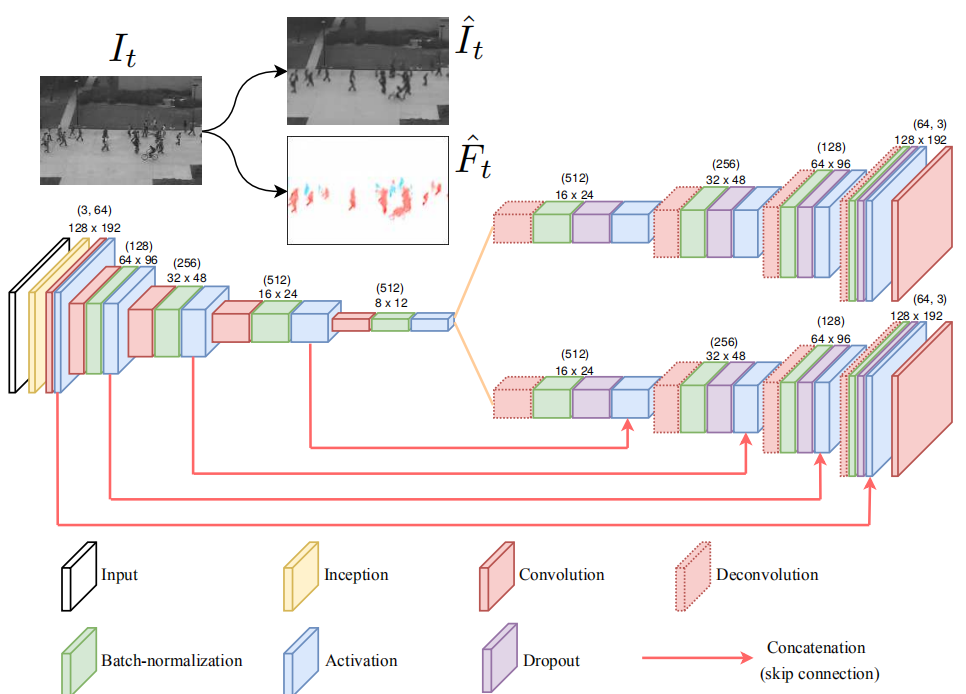
A drawback of using only loss is the blur in the output. They add a constraint to preserve the original gradient (sharpness) in the reconstructed image.

Final loss:

Motion Prediction U-Net:

U-Net focuses on learning the association between normal patterns and corresponding motions.

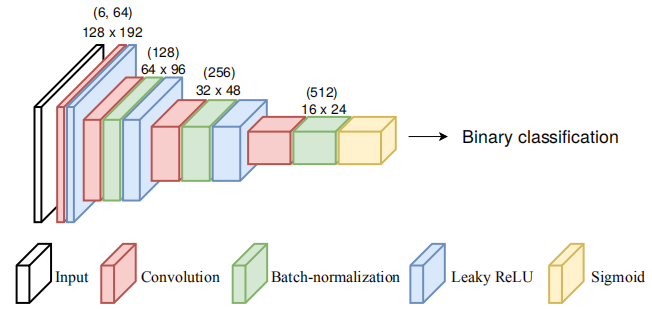
The ground truth optical flow is estimated by FlowNet2.



Additional Motion-related Objective Function:

The classification is performed according to image patches instead of the whole image. They constrain patches at feature-level so that each feature map must attempt to provide a classification result.

Where x, y and c indicate the spatial position and the corresponding channel of a unit in the feature maps outputted from D.



Anomaly Detection:

An anomalous event occurring within a small image region may be missed due to the summation and/or average operations over all pixel positions. They propose another score estimation scheme considering only a small patch instead of the entire frame.

First, define partial scores individually estimated on the two model streams sharing the same patch position:

Where P indicates an image patch and |P| is its number of pixels.

Frame-level score is computed as a weighted combination of two partial scores:

Where and are the weights calculated according to training data. is to control the contribution of partial scores to the summation. is the patch providing the highest value of in the considering frame.

The wights and are estimated as the inverse of average scores obtained on the training data of n images:

The size of P was set to 16 x 16. Such patches are determined by a sliding window. since the model focuses on motion prediction efficiency.

Final frame-level score:

Where t is the frame index in a video containing m frames.