where n is the prediction error history that influences the anomaly detection module's internal model for detecting event boundaries. In our experiments, we maintain n at 5. This is chosen based on the average response time of human perception, which is around 200 ms [33].

The gating signal, $G(t)$, is triggered when the current prediction error exceeds the average quality metric by at least 50%.

$$G(t) = \begin{cases} 1, & \frac{E_P(t)}{P_q(t-1)} > \psi_e \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $P_E(t)$ is the perceptual prediction error at time t, $G(t)$ is the value of the gating signal at time t, $P_q(t-1)$ is the prediction quality metric at time t and Ψ_e is the prediction error threshold for boundary detection. For optimal prediction, the perceptual prediction error would be very high at the event boundary frames and very low at all within-event frames. In our experiments, Ψ_e is set to be 1.5.

In actual, real-world video frames, however, there exist additional noise in the form of occlusions and background motion which can cause some event boundaries to have a low perceptual prediction error. In that case, however, the gating signal would continue to be low and become high when there is a transient increase in error. This is visualized in Figure 1. It can be seen that the perceptual errors were lower at event boundaries between activities *take bowl* and *crack eggs* in a video of ground truth *make pancakes*. However, the prediction error increases radically soon after the boundary frames, indicating a new event. Such cases could, arguably, be attributed to conditions when there are lesser variations in the visual features at an event boundary.

3.4. Adaptive Learning for Plasticity

The proposed training of the prediction module is particularly conducive towards overfitting since we propagate the perceptual prediction error at every time step. This introduces severe overfitting, especially in the prediction model. To allow for some plasticity and avoid catastrophic forgetting in the network, we introduce the concept of adaptive learning. This is similar to the learning rate schedule, a commonly used technique for training deep neural networks. However, instead of using predetermined intervals for changing the learning rates, we propose the use of the gating signal to modulate the learning rate. For example, when the perceptual prediction rate is lower than the average prediction rate, the predictor model is considered to have a good, stable representation of the current event. Propagating the prediction error, when there is a good representation of the event can lead to overfitting of the predictor model to that particular event and does not help generalize. Hence, we propose lower learning rates for time steps when there are negligible prediction error and a relatively higher (by a magnitude of 100) for when there is higher prediction error. Intuitively, this adaptive learning rate allows the model to adapt much quicker to new events (at event boundaries where there are likely to be higher errors) and learn to maintain the internal representation for within-event frames.

Formally, the learning rate is defined as the result of the adaptive learning rule defined as a function of the perceptual prediction error defined in Section 3.2 and is defined as

$$\lambda_{learn} = \begin{cases} \Delta_t^- \lambda_{init}, & E_P(t) > \mu_e \\ \Delta_t^+ \lambda_{init}, & E_P(t) < \mu_e \\ \lambda_{init}, & otherwise \end{cases} \tag{7}$$

where Δ_t^-, Δ_t^+ and λ_{init} refer to the scaling of the learning rate in the negative direction, positive direction and the initial learning rate respectively and $\mu_e = \frac{1}{t_2-t_1}\int_{t_1}^{t_2} E_P \, dE_P$. The learning rate is adjusted based on the quality of the predictions characterized by the perceptual prediction error between a temporal sequence between times t_1 and t_2, typically defined by the gating signal.. The impact of the adaptive changes to the learning rate is shown in the quantitative evaluation Section 4.4, where the adaptive learning scheme shows improvement of up to 20% compared to training without the learning scheme.

3.5. Implementation Details

In our experiments, we use a VGG-16 [31] network pre-trained on ImageNet as our hierarchical, feature encoder module. We discard the final layer and use the second fully connected layer with 4096 units as our encoded feature vector for a given frame. The feature vector is then consumed by a predictor model. We trained two versions, one with an RNN and the other with an LSTM as our predictor models. The LSTM model used is the original version proposed by [15]. Finally, the anomaly detection module runs an average low pass filter described in Section 3.3. The initial learning rate described in Section 3.4 is set to be 1×10^{-6}. The scaling factors Δ_t^- and Δ_t^+ are set to be 1×10^{-2} and 1×10^{-3}, respectively. The training was done on a computer with one Titan X Pascal.

4. Experimental Evaluation

4.1. Datasets

We evaluate and analyze the performance of the proposed approach on three large, publicly available datasets - Breakfast Actions [19], INRIA Instructional Videos dataset[3] and the 50 Salads dataset [32]. Each dataset offers a different challenge to the approach allow us to evaluate its performance on a variety of challenging conditions.

Breakfast Actions Dataset is a large collection of 1,712 videos of 10 breakfast activities performed by 52 actors.