**Motivation**

- Most of the multi-modal video recognition methods are computationally expensive, as they usually process all the data (including redundant/irrelevant parts).
- Utilizing information from all the input modalities may be counterproductive as informative modalities are often overwhelmed by uninformative ones in long videos.
- Some modalities require more computation than others and hence selecting the cheaper modality with good performance can significantly save computation leading to more efficient recognition.

**Key Idea**

- AdaMML: A novel and differentiable approach to learn a decision policy that selects optimal modalities conditioned on the inputs for efficient video recognition.
- This is in sharp contrast to current multi-modal learning approaches that utilizes all the input modalities without considering their relevance to the video recognition.
- Learn a lightweight model (referred to as the multi-modal learning approaches) that outputs the posterior probabilities of all the binary relevance to the video recognition.
- Some modalities require more computation than others and hence selecting the cheaper modality with good performance can significantly save computation leading to more efficient recognition.

**Framework**

- AdaMML consists of a policy network and a recognition network composed of different sub-networks that are trained jointly (via late fusion with learnable weights) for recognizing videos.
- Policy network decides what modalities to use per segment to achieve the best accuracy and efficiency in video recognition.
- In training, policies are sampled from a Gumbel-Softmax distribution, which allows us to optimize the policy network via standard backpropagation.
- During inference, an input segment is first fed into policy network and then selected modalities are routed to recognition network to generate segment-level predictions. Finally, the network averages all segment-level predictions to obtain video-level prediction.

**Multi-Modal Policy Network**

The policy network contains a lightweight joint feature extractor and an LSTM module for modeling the causality across different time steps in a video.

\[ h_t, a_t = \text{LSTM}(f_t, h_{t-1}, a_{t-1}) \]

Given the hidden state, the policy network estimates a policy distribution for each and then samples binary decisions indicating whether to select a modality at time step via Gumbel-Softmax operation.

**Training using Gumbel-Softmax Sampling**

An efficient way to replace the non-differentiable sample from a discrete distribution with a differentiable sample from a corresponding Gumbel-Softmax distribution.

\[ P_t = \text{argmax}(\log z_{t,k} + G_{t,k}), \quad k \in [1, \ldots, K] \]

\[ P_t = \exp(\log z_{t,k} + G_{t,k}) / \sum_{i=1}^{K} \exp(\log z_{t,i} + G_{t,i}) \]

During forward pass, we sample the policy and during the backward pass, we approximate the gradient of the discrete samples by computing the gradient of the continuous softmax relaxation.

**Loss Function**

\[ E_{P_t, \alpha_t} \rightarrow -p \log(P(V, A)) - \sum_{t \in T} \log C_t \] if correct otherwise

First part: standard cross-entropy loss to measure the classification quality; Second part: drives the network to learn a policy that favors selection of modality that is computationally more efficient in recognizing videos.

**Results**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Kinetics 400</th>
<th>ActivityNet 200</th>
<th>ActivityNet 1206</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB</td>
<td>45.6</td>
<td>58.5</td>
<td>70.0</td>
</tr>
<tr>
<td>Audio</td>
<td>40.9</td>
<td>52.7</td>
<td>64.0</td>
</tr>
<tr>
<td>Weighted Fusion</td>
<td>46.5</td>
<td>58.5</td>
<td>71.0</td>
</tr>
</tbody>
</table>

Consistently outperforms all hand-designed fusion strategies (~50% FLOPS savings) without having any accuracy penalties.

**Visualizations**

Qualitative examples showing the effectiveness of AdaMML in selecting the right modalities per video segment (marked by green borders). Our adaptive approach focuses on right modalities to use per segment for correctly classifying videos while taking efficiency into account.