Introduction

Goal: Better describe video events with natural language using past, present, and future events within a given video.

Applications: Video query search, ad content matching, video multimedia editing, security, and more.

Dataset


Method & Model

Baseline Attention Module:
- Max pooling applied to all C3D-PCA feature vectors
- Output = concatenation of pooled representations of past, present, and future

\[ h_i^{\text{past}} = \frac{1}{Z_i^{\text{past}}} \sum_{j \neq i} 1\{f_j^{\text{end}} < f_i^{\text{end}}\} a_{ij} h_j \]

Where we have

\[ Z_i^{\text{past}} = \sum_{j \neq i} 1\{f_j^{\text{end}} < f_i^{\text{end}}\} \]

\[ w_i = w_a h_i + b_a \]

\[ a_{ij} = w_i h_j \]

Future features are analogous Vectorized for efficient computation

NetVLAD (Learnable Pooling):

- Computes clusters in the input features and residuals from the input features to the cluster center, then multiplies to a softmax

\[ VLAD(j, k) = \sum_i^{N} \text{softmax}(h_i)(h_{ij} - c_{k,j}) \]

(nonlinear relationship between C3D frame features)

Context Gating (CG)

\[ CG(X) = \sigma(WX + B) \ast X \]

(quadratic relationship between NetVLAD output features X)

Captioning Module:
- Concatenation of Word Embeddings and CG output as input to 2-layer LSTM. Each step of LSTM has identical values for proposal activations with the corresponding word embedding for that time step.

Caption Generation (test):
- Greedy search
- Sampling
- Beam search

Loss Function & Optimization:
- Cross Entropy
- Gradient Descent with Momentum
- Adam Optimizer

Results & Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU Mean Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.542 0.533 0.554 0.585</td>
</tr>
<tr>
<td>NetVLAD</td>
<td>0.607 0.589 0.592 0.598</td>
</tr>
</tbody>
</table>

Future Work

- GPU support from Adobe to train on 20k videos
- Combine strengths of NetVLAD and baseline
- Evaluate captions with METEOR and CIDEr metrics

Attention Weights:

- GPU support from Adobe to train on 20k videos
- Combine strengths of NetVLAD and baseline
- Evaluate captions with METEOR and CIDEr metrics