Video Summarization using Transformers

Manjot Bilkhu, Siyang Wang, Tushar Dobhal
Video Summarization

Input: A piece of video
Output: Text description of the video

“A man is cooking.”

“A man is cooking food. He is putting lemon on it. The food is on a plate.”
Datasets

1. MSVD
   a. 1,970 YouTube video clips with 10 to 25 seconds videos
   b. Each clip is annotated with sentences in multiple languages
   c. Consists of about 41 per video and 85k English sentences in total

2. ActivityNet
   a. 203 activity categories with an average of 137 untrimmed videos per class and 1.41 activity instances per video, making a total of 20k videos totaling 849 hours
   b. Long videos, descriptions in paragraphs rather than sentences
   c. Aimed at activity recognition (a vanilla classification task)
   d. Can be used for ‘Dense’ captioning
Video Feature Extraction

- **Spatial Features** -
  - ImageNet pre-trained VGG and ResNet models
  - Captures spatial features from RGB
  - Requires Optical flow images for capturing temporal features

- **Spatio-temporal Features** -
  - Sports-1M/ UCF-101/ Kinetics pre-trained C3D and I3D models
  - Captures spatial and temporal features from the video directly using RGB frames
  - Optical flow level features help improve motion feature capture

- Both of the above are actively used for Video Captioning with 3D features often giving better encoding of the video
C3D Networks

Figure 1. **2D and 3D convolution operations.** a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.

Figure 3. **C3D architecture.** C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from pool1 to pool5. All pooling kernels are $2 \times 2 \times 2$, except for pool1 is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.
I3D : Inflated 3D Convolutions

- Inflate 2D ConvNets into 3D
- Bootstrap 3D filters from 2D Filters (Imagenet)
- Two 3D Streams
- Converted Inception Network to 3D
Transformers

- Removed Encoder Embedding Layer
- Used \( d_{\text{model}} = 500 \) for ActivityNet
- Used 10 Attention Heads
- Altered Learning Rate and its scheduling
  - No Warm-up steps: Blows up training
Universal Transformers

- Removed Encoder Embedding layer
- Used 8 layers instead of 4
- Used $d_{\text{model}} = 500$ instead of 512
- Used 10 attention heads
- Altered Learning Rate and its scheduling
Related Work

- Pan et al. LSTM-TSA (transferred semantic attributes) used an LSTM network for caption prediction
  - RGB frames (VGG) and 3-D video features (C3D) serve as input
  - Used text features as well from another CNN network

- Long et al. Keyless attention mechanism
  - RGB and Audio inputs to their Bi-directional LSTM network
  - Used RGB input images with audio input
  - Used a 2-layer feedforward network as attention mechanism

- Long et al. Multi-faceted attention for text, video and RGB
  - Applied attention to each input stream
  - Used a separate network for text features

- All the above techniques generate single sentence captions
Related Work contd...

- End-to-End Dense Captioning (Socher et al.) used a network with one encoder and two decoders based on vanilla transformers.
- Purpose is to identify important segments in the video for dense video captioning.
- Used RGB and Optical flow features from a ResNet-152 as input.
- First demonstrated use of a transformer based model for video captioning.
- Generated one sentence per segment of the video. On average 3-4 per video.
Architecture

- Pre-trained C3D on UCF-101 dataset
  - 4096 features
  - Dimensionality Reduction using PCA
  - 512 features

- Pre-trained RGB-I3D on Kinetics dataset
  - 1024 features
  - Dimensionality Reduction using PCA
  - 512 features

- Pre-trained Flow-I3D on Kinetics dataset
  - 1024 features
  - Dimensionality Reduction using PCA
  - 512 features
## Results

### MSVD

<table>
<thead>
<tr>
<th>Model</th>
<th>BLUE-1</th>
<th>BLUE-2</th>
<th>BLUE-3</th>
<th>BLUE-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-TSA</td>
<td>0.828</td>
<td>0.72</td>
<td>0.628</td>
<td>0.528</td>
</tr>
<tr>
<td>Multi-faceted Attention</td>
<td>0.830</td>
<td>0.719</td>
<td>0.630</td>
<td>0.520</td>
</tr>
<tr>
<td><strong>C3D + Transformer (ours)</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.755</strong></td>
<td><strong>0.607</strong></td>
<td><strong>0.516</strong></td>
</tr>
<tr>
<td><strong>I3D + Transformer (ours)</strong></td>
<td><strong>0.89</strong></td>
<td><strong>0.73</strong></td>
<td><strong>0.564</strong></td>
<td><strong>0.442</strong></td>
</tr>
<tr>
<td><strong>C3D + Universal Transformer (ours)</strong></td>
<td><strong>0.901</strong></td>
<td><strong>0.765</strong></td>
<td><strong>0.587</strong></td>
<td><strong>0.501</strong></td>
</tr>
<tr>
<td><strong>I3D + Universal Transformer (ours)</strong></td>
<td><strong>0.91</strong></td>
<td><strong>0.782</strong></td>
<td><strong>0.521</strong></td>
<td><strong>0.46</strong></td>
</tr>
</tbody>
</table>
Analysis: The Good...

Model: 'a man is playing a guitar'

Ground Truth: ['a guy is playing guitar', 'a man is playing a guitar', 'a man is playing the guitar']
Model: 'a man is pouring oil to a pan'

Ground Truth: GT:- ['a man' is adding oil to a pan', 'A man is pouring oil into a pan', 'Someone is pouring oil into a pot']
Analysis: The Good...

Model: 'men are playing table tennis'

Ground Truth: ['Men are playing ping pong', 'two men are playing table tennis']
Analysis: The Bad...

Model: 'a cat is walking'

Ground Truth: ['A turtle is walking', 'the tortoise is moving']
Analysis: The Bad...

Model: 'men are fighting'

Ground Truth: ['Two men are practicing karate', 'Two men are boxing']
Analysis: The Ugly...

Model: 'a man is playing with a dog'

Ground Truth: ['A body builder is doing exercises', 'A bodybuilder is doing exercise']
Model: 'a gymnast jumps onto a gymnastics beam <.> he does a gymnastics routine on the beam beam <.> he then goes to lands on the mat <.> '
Model: 'A woman is seen speaking to the camera while holding a animated object <.> she then plays playing the violin while the man continues to play and skips and ends to lay up the point <.>.'
Analysis: Dense Captioning: And The Bad...

Model: 'a cowboy is standing on the back in a field hats riding a horse around there of down open a man of horses appear life hand, and they as lifts hands all horns waiting several jumps'
Future Goals

- **Short-term Goals:**
  - Get BLEU score statistics for Activity Net (still under training)
  - Correct the problems with nouns that the model is facing. Any suggestions are welcomed

- **Long-term Goals:**
  - I3D is pretrained on Kinetics-400 and C3D on Sports-101. Fine tune C3D and I3D to our datasets. It has shown to improve performance
  - Modify ACT or come up with some way to improve the halting process in Universal Transformer
Thank You