Supplementary Materials for ICCV 2017

**Attention-Based Multimodal Fusion for Video Description**

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham,

John R. Hershey, Tim K. Marks, Kazuhiro Sumi

Mitsubishi Electric Research Laboratories (MERL)

Evaluation results on the YouTube2Text test set. *Unimodal (VGG-16)* indicates a method that uses only VGG-16 image features (no spatiotemporal features nor audio features). *Naïve Fusion* is the conventional approach using temporal attention only (see Figure 3 in the paper). *Attentional Fusion* is our proposed Multimodal Attention approach (see Figure 4 in the paper). The symbol (V) denotes methods that use only the two visual modalities: VGG-16 image features and C3D spatiotemporal features. The symbol (AV) denotes our methods that use all three modalities (audio features as well as the two types of video features). The last column shows the CIDEr scores obtained sentence by sentence. To see each video, please click or Ctrl+click on the image or URL.

|  |  |  |  |
| --- | --- | --- | --- |
| <nULE40HEWpA_5_11.mp4> | Attention | Description | CIDEr |
| Unimodal (VGG-16) | a man is playing | 0.386 |
| Naïve Fusion (V) | a person is playing an acoustic keyboard | 0.153 |
| **Attentional Fusion (V)** | **a kitten is playing** | **2.585** |
| Discussion | Our Attentional Fusion (V) method worked best. *Note:* This is a silent video (there is no audio). | | |

|  |  |  |  |
| --- | --- | --- | --- |
| <t4aPGtx7e6k_0_10.mp4> | Attention | Description | CIDEr |
| Unimodal (VGG-16) | a monkey is running | 1.597 |
| Naïve Fusion (V) | a dog is playing | 0.765 |
| Naïve Fusion (AV) | a monkey is running | 1.597 |
| **Attentional Fusion (V)** | **a monkey is pulling a dogs tail** | **1.658** |
| **Attentional Fusion (AV)** | **a monkey is playing** | **2.222** |
| Discussion | Our Attentional Fusion methods worked best. Attentional Fusion (AV) is ranked best by CIDEr (based on the ground-truth labels that were generated by Amazon Mechanical Turk workers). Attentional Fusion (V) is also boldfaced here because although it has a lower CIDEr score, we feel it is a more accurate description of the video. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| <m1NR0uNNs5Y_73_78.mp4> | Attention | Description | CIDEr |
| Unimodal (VGG-16) | a man is slicing a potato | 0.079 |
| Naïve Fusion (V) | a woman is cutting an onion | 1.560 |
| **Naïve Fusion (AV)** | **a woman is peeling an onion** | **4.272** |
| Attentional Fusion (V) | a man is slicing a potato | 0.079 |
| **Attentional Fusion (AV)** | **a woman is peeling an onion** | **4.272** |
| Discussion | Our inclusion of audio features (AV methods) enabled the network to identify the “peeling” action. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| <yU5sxW9bErQ_0_17.mp4> | Attention | Description | CIDEr |
| Unimodal (VGG-16) | a woman is riding a horse | 4.517 |
| Naïve Fusion (V) | a girl is riding a horse | 4.380 |
| Naïve Fusion (AV) | a girl is riding a horse | 4.380 |
| **Attentional Fusion (V)** | **a man is riding a horse** | **5.805** |
| Attentional Fusion (AV) | a girl is riding a horse | 4.380 |
| Discussion | Our Attentional Fusion (V) method worked best. Adding audio features hurt performance because the audio track contains only overdubbed music. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| <kk3TIio1-Uw_5_14.mp4> | Attention | Description | CIDEr |
| Unimodal (VGG-16) | a man is singing | 0.071 |
| Naïve Fusion (V) | a man is singing | 0.071 |
| Naïve Fusion (AV) | a man is playing a guitar | 1.483 |
| Attentional Fusion (V) | a man is playing a guitar | 1.483 |
| **Attentional Fusion (AV)** | **a man is playing a violin** | **2.276** |
| Discussion | Our Attentional Fusion (AV) method worked best. Individually, audio features and multimodal attention each improve performance, but both must be combined to correctly identify “violin.” | | |