Stanford University Dense-Captioning Videos with Learnable Pooling Stanford University

Songze Li, Julio A. Martinez, Parker Miller

Introduction

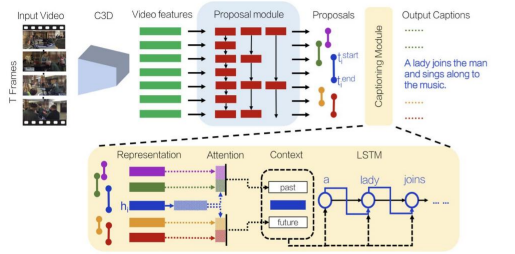
Goal: Better **describe video events with natural language** using past, present, and future events within a given video.

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

Dataset

Random sample of videos from **ActivityNet Captions**: **2000 train**, **250 validation**, and **250 test**.

Method & Model



Baseline Attention Module:

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

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

Where we have

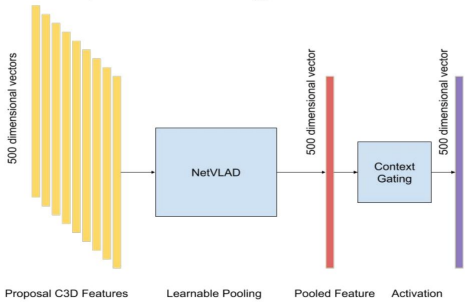
$$Z_i^{past} = \sum_{j \neq i} 1\{f_j^{end} < f_i^{end}\}$$

$$w_i = w_a h_i + b_a$$

$$a_{ij} = w_i h_j$$

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**

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

(nonlinear relationship between C3D frame features)

Context Gating (CG)

$$CG(X) = \sigma(WX + B) * X$$

(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

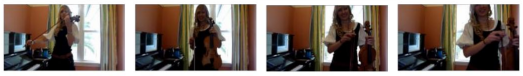
Loss Function & Optimization:

- Cross Entropy
- Gradient Descent with Momentum
- Adam Optimizer

Results & Analysis

BLEU Mean Values				
Model	**B1**	**B2**	**B3**	**B4**
Baseline	**0.542**	**0.533**	**0.554**	**0.585**
NetVLAD	**0.607**	**0.589**	**0.592**	**0.598**

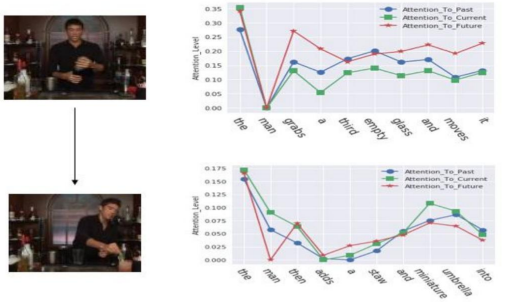
Ground Truth: another man is giving instruction on lawn and garden care <end>
Baseline: he goes down giving slide on lawn and garden care little another <end>
NetVLAD: a man is shown wandering through the yard <end>



Ground Truth: the woman stops playing and smiles once more before walking out of view of the camera putting the violin down and ending the song <end>
Baseline: the woman stops playing and smiles once more before walking walking of of view the the putting putting violin violin violin ending ending the the the the the the
NetVLAD: the woman stops playing and smiles once down into the camera putting the violin down and ending the song <end>

Ground Truth: he also has protective gloves on that are brown in color <end>
Baseline: he also has protective gloves on that are brown in color <end>
NetVLAD: he has protective gloves on that are brown <end>

Attention Weights:



Future Work

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