Localizing Moments in Video with Natural Language

11775 – Midterm Progress

Presentation

Anusha Prakash Anuva Agarwal

Bhavya Karki Pravalika Avvaru

Date: March 28th 2018

Carnegie Mellon University
School of Computer Science

Language Technologies Institute

Outline

Carnegie Mellon

- Introduction
- Related Work
- Dataset
- Baseline Architecture
- Initial Experimental Results
- Proposed methodologies / Next Steps
- Logistics
- References
- Q&A

Introduction

Introduction

Carnegie Mellon

Baby twitching after eating lemon

Shaking head vigorously

Rubbing eyes after eating

Baby squinting eyes on being offered the lemon

Baby licking lime the third time

Baby laughing

Feeding a baby

Pushing food away when fed again

When does a particular activity occur in a video ?

Problem statement: **Retrieve a specific temporal segment, or moment from a video given a natural language text description.**

- Text assisted Video Editing
- Video Retrieval
- Finding moments in long video footages
- Finding moments from a personal holiday video
- B-roll stock video footages from a large video library (Shutterstock, Adobe)

Related Literature

Carnegie Mellon

1. **Generation and Comprehension of Unambiguous Object Descriptions** **[Junhua et al, 2016]**
   a. Generate unambiguous description of a specific object + comprehend or interpret such an expression
   b. Present a new large scale dataset for referring expressions based on MS COCO

2. **Video Object Segmentation with Language Referring Expressions** **[Anna et al, 2018]**
   a. High quality video object segmentation results using language referring expressions
   b. Performs on par with semi-supervised methods with access to the pixel-accurate object mask.
   c. Evaluated on DAVIS'17 dataset

3. **Modeling Relationships in Referential Expressions with Compositional Modular Networks** **[Ronghang et al, 2016]**
   a. Compositional Modular Networks (CMNs) - learn language representation and image region localization jointly
   b. Two types of modules - (i) localizing specific textual components by outputting unary scores (ii) relationship between two pairs of bounding boxes by outputting pairwise scores

Related Literature

Carnegie Mellon

4. **Learning Joint Representations of Videos and Sentences with Web Image Search** **[Otani et al, 2016]**

    a. Web image search in sentence embedding process to disambiguate fine-grained visual concepts
    b. Embedding models for multimodal inputs whose parameters are learned simultaneously

5. **Deep fragment embeddings for bidirectional image sentence mapping** **[Karpathy et al, 2015]**

    a. Embeds fragments of images and sentences into a common space.
    b. Retrieve relevant images given a sentence query, + relevant sentences given an image query
    c. Stanford CoreNLP parser to compute the dependency trees for every sentence.
    d. Evaluated on Flickr8k and Flickr30k datasets

Dataset (DiDeMo)

Carnegie Mellon

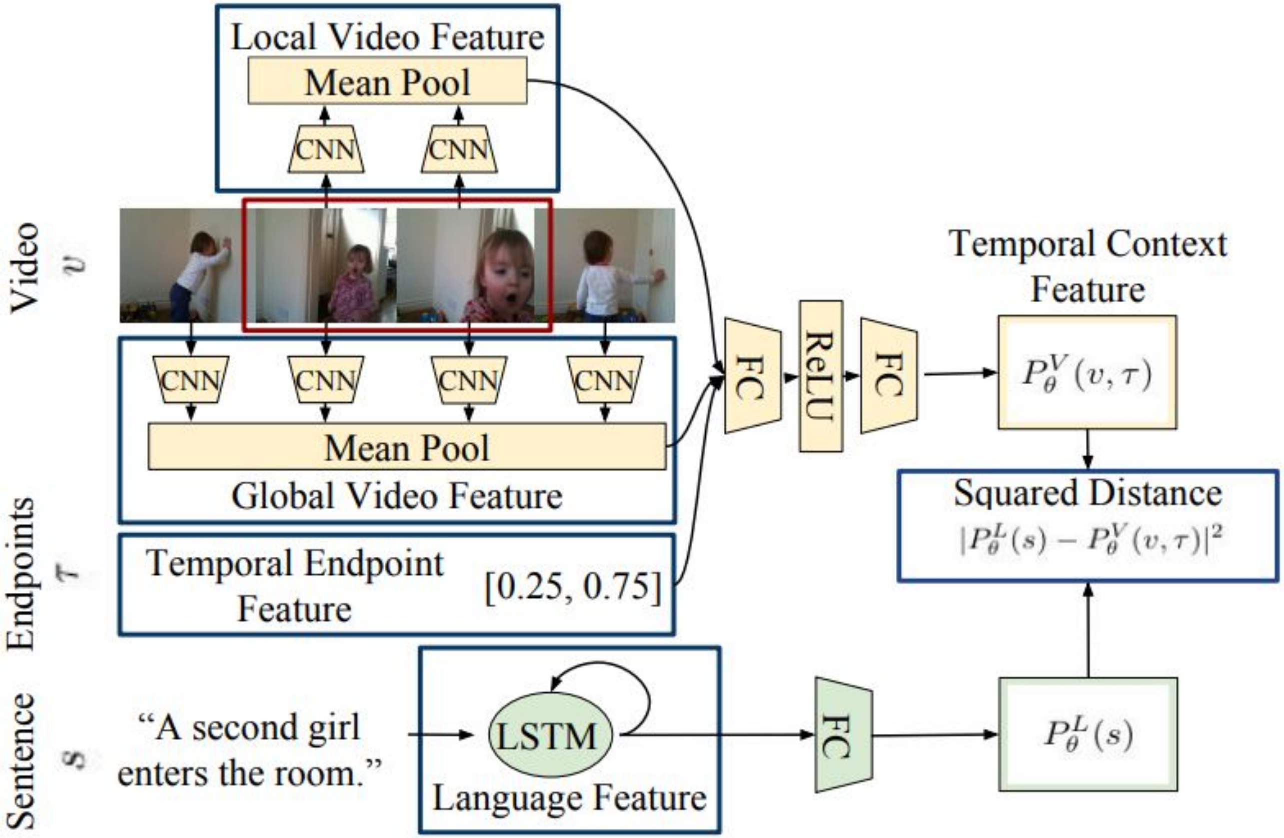
- Distinct Describable Moments (DiDeMo)
- Dataset consists over 10,000 videos
- 25-30 seconds long personal videos randomly selected from Flickr
- Over 40,000 localized text descriptions (3-5 pairs per video)
- Represent a diverse set of real-world videos like pets, concerts, sports
- Higher percentage of temporal indicators, spatial indicators and verbs
- Consist of both eventful and uneventful segments in the video

Dataset (DiDeMo)

Carnegie Mellon


```
{
    "num_segments": 6,
    "description": "the toddler puts her head on the ground.",
    "dl_link": "https://www.flickr.com/video_download.gne?id=3926817284",
    "times": [
        [
            3,
            3
        ],
        [
            2,
            3
        ],
        [
            3,
            3
        ],
        [
            3,
            4
        ]
    ],
    "video": "75319260@N00_3926817284_e685e53cef.3gp",
    "annotation_id": 8213
},
```

Baseline Architecture - Moment Context Network (MCN)



[Hendricks et al, 2017]

Baseline Architecture contd..

- **Model (MCN):**
  - Joint Video-Language Model - Shared Embedding Space
  - Glove Embedding
  - LSTM
  - CNN Layers (Local and Global)
  - Fully Connected Layer
  - Ranking Loss
  - Distance Measures (Euclidean)

- **Features:**
  - Temporal Endpoint Features - When a moment occurs in a video
  - Low level
    - Optical flow
  - High level
    - RGB - VGG Net FC7
  - Global Video Features - Provides Temporal Context
  - Late Fusion

Evaluation Metrics

Carnegie Mellon

- Rank@1
- Rank@5
- Mean Intersection over Union (mIoU)
- **Baseline: Moment Frequency Prior** - Tendency to select short moments towards the beginning of videos. It selects moments which correspond to gifs most frequently described by annotators.

Model	Rank@1	Rank@5	mIoU
Moment Frequency Prior	**19.40**	**66.38**	**26.65**

Baseline Experiments

Carnegie Mellon

Lambda - 0.5

No_of_epochs - 30,000

Features - LSTM-Fusion + global + tef (MCN)

Model	Average IOU	AverageRank@1	AverageRank@5
Reproduced Baseline	**0.405315**	**0.270828**	**0.785377**
Baseline	0.4108	0.2810	0.7821

Initial Experiments (Glove6B)

Carnegie Mellon

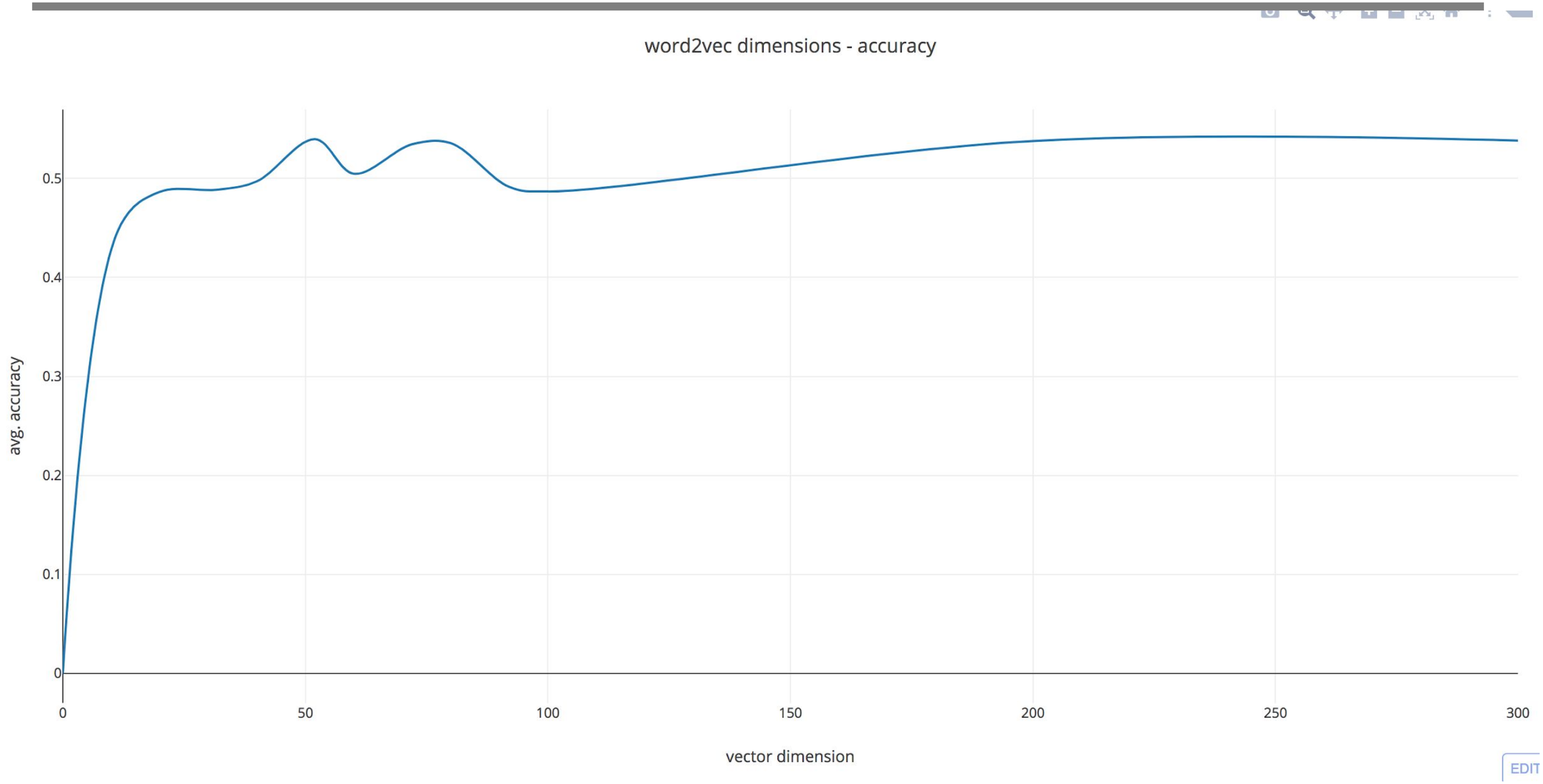
Lambda - 0.5

No_of_epochs - 10,000

Features - LSTM-Fusion + global + tef (MCN)

Model	Average IOU	AverageRank@1	AverageRank@5
200 word dimensional embeddings	**0.389062**	**0.2628**	**0.762746**
300 word dimensional embeddings (Baseline)	0.386642	0.261378	0.772196

Initial Experiments (Glove6B)



Initial Experiments (Language Model)

Carnegie Mellon

Lambda - 0.5

No_of_epochs - 10,000

Features - LSTM-Fusion + global + tef (MCN)

Model	Average IOU	AverageRank@1	AverageRank@5
RNN for Language	**0.253292**	**0.191246**	**0.262373**
LSTM for Language	0.386642	0.261378	0.772196

Proposed Methodologies / Next Steps

- Ablation study with word2vec and various glove embedding for the language model network initialization
- Implement Bi-LSTM, GRU and Hierarchical RNN approaches to the language language model
- Explore better and different distance metrics to build the joint-embedding space of the video and language
- Experiment with Early and Double Fusion of the visual features

Proposed Methodologies / Next Steps

Carnegie Mellon

- Bilinear transforms - using Bi(symmetrical) DNNs
- Use features for all moments instead of just 6 moments (stride)
- Extract richer local and global visual features and employ a Bi-LSTM to combine these to produce temporal context features
- To address up-scaling the vocabulary part, we plan to pre-train on Moments in Time dataset and select relevant actions and find a common embedding space.

Infrastructure / Logistics

Carnegie Mellon

- We have been provided with AWS credits and we have access to the PSC cluster
- Our initial experiments were run on p2.xLarge (1 quad-core, GPU instance with 61 GB that is priced at $0.900 per hour)
- Our models were trained for 10,000 epochs and each model takes around an hour
- Testing the model takes around 20 minutes

References

- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. CoRR , abs/1511.02283, 2015.

- Anna Khoreva, Anna Rohrbach, Bernt Schiele. Video Object Segmentation with Language Referring Expressions. CVPR, (arXiv:1803.08006), 2018.

- Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. CoRR , abs/1611.09978, 2016.

- Andrej Karpathy, Armand Joulin, and Fei-Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. CoRR , abs/1406.5679, 2014.

- Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. CoRR , abs/1608.02367, 2016.

Q&A