Localizing Moments in Video with Natural Language

Anusha Prakash Anuva Agarwal

Bhavya Karki Pravalika Avvaru

Date : May 2nd 2018

Carnegie Mellon University
School of Computer Science

Language Technologies Institute

Outline

Carnegie Mellon

- Problem statement
- Recap
- Proposal
- Experiments and Analysis
- Important Takeaways
- References
- Q & A

PROBLEM STATEMENT

Carnegie Mellon

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.**

RECAP….. (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

RECAP……(Videos + Queries)

Carnegie Mellon

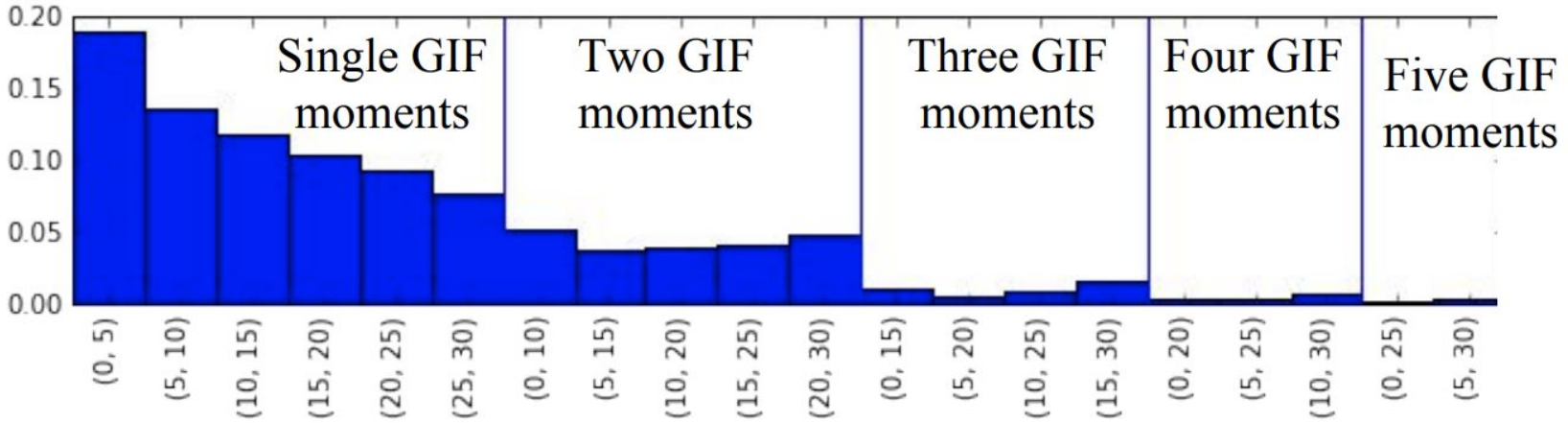
A girl stands up

Camera zooms in

White car passes by

Data Statistics

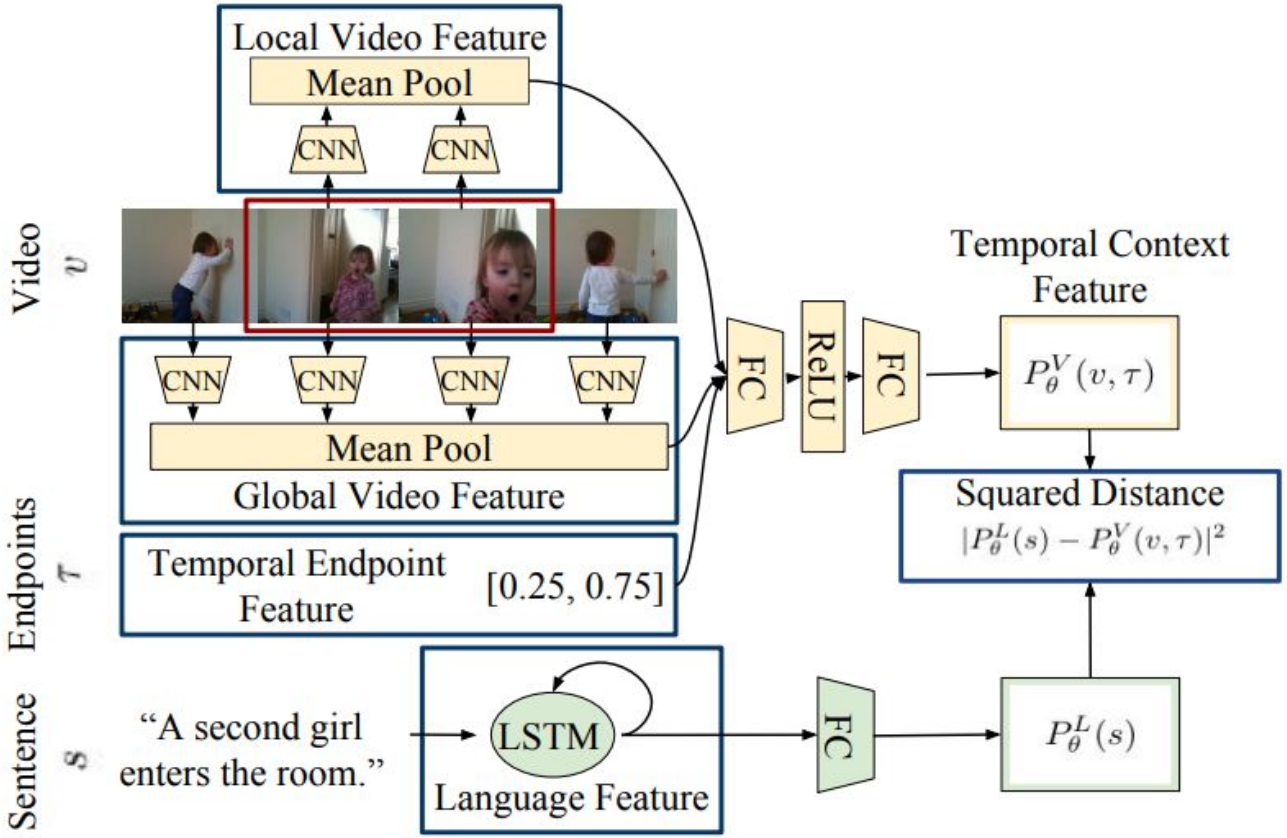
Carnegie Mellon



Distribution of segments marked in DiDeMo. Moments tend to be short and occur towards the beginning of videos

RECAP……(Baseline Architecture)

Carnegie Mellon



[Hendricks et al, 2017]

RECAP…..(Baseline Architecture)

Carnegie Mellon

- **Model (Moment Conext Network):**
  - 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

RECAP……(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**

Proposal

Carnegie Mellon

- ✓ Baseline
- ✓ Language Embedding
- ✓ Language Model
- ✓ Joint-embedding space distance metrics
- ✓ Visual Features
- ✓ Fusion Techniques
- ✓ Bilinear transforms - using Bi(symmetrical) DNNs
- ✓ To address up-scaling the vocabulary part, we planned to pre-train on ~~Moments in Time dataset~~ and select relevant actions.
- ✓ Audio Features - SoundNet
- ✓ I3D Activity Recognition
- ✓ Image Captioning

Kinetics

Experiments And Analysis

Baseline Experiments

Carnegie Mellon

Lambda - 0.5

No_of_epochs - 30,000

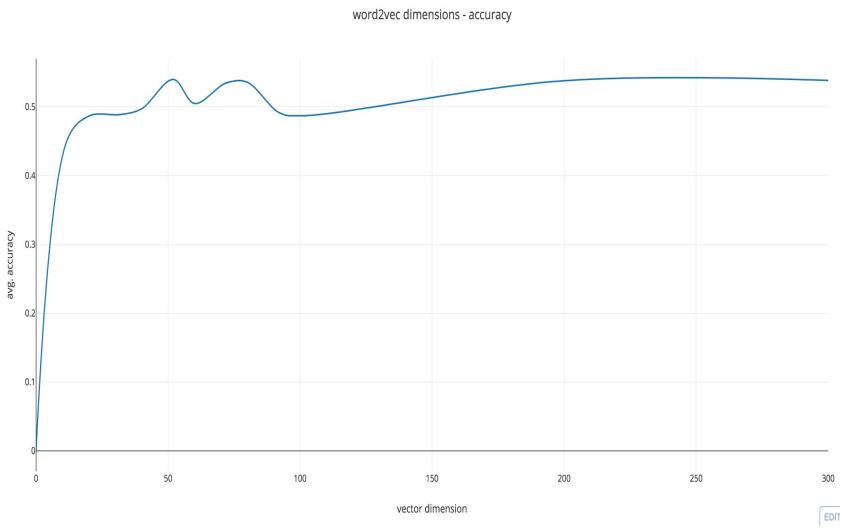
Features - LSTM-Fusion + global + tef (MCN)

Model	Average IOU	AverageRank@1	AverageRank@5
Reproduced Baseline	**0.407195**	**0.277861**	**0.781841**
Baseline	0.4108	0.2810	0.7821

Language Embedding Space

Carnegie Mellon

- **Approach** : 200d GloVE 6B instead of 300d GloVE 6B
- **Motivation** : Vocab small, compact embedding than sparse embedding
- **Results** : 200d GloVE 6B performs slightly better than 300d GloVE 6B



Model	Avg IOU	Avg Rank@1	Avg Rank@5
200 D	**0.389062**	**0.2628**	**0.762746**
300 D	0.386642	0.261378	0.772196

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**
BiLSTM for Language	**0.238544**	**0.285571**	**0.723228**
LSTM for Language	0.386642	0.261378	0.772196

Why LSTM is better than the rest ??

Carnegie Mellon

RNN - Vanishing gradients, does not learn sequences well

BiLSTM - Reversing of inputs to LSTM layers hinders sequence of moments - loss of temporal context

LSTM: Handles problems incurred by both above models

Example:

a person holds up two fingers
a man is counting on four fingers
man counts two **again**

Common Embedding Space - Distance Metrics

Carnegie Mellon

Lambda - 0.5

No_of_epochs - 10,000

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

Model	Avg IOU	AvgRank@1	AvgRank@5
Dot product	**0.39201**	**0.26719**	**0.78121**
Element-wise	**0.373324**	**0.26013**	**0.73450**
Euclidean	0.386642	0.261378	0.772196

Feature Comparisons

Carnegie Mellon

A dog looks at the camera and jumps at it.

RGB

Flow

Fusion

Brown dog runs at the camera.

local

local+global

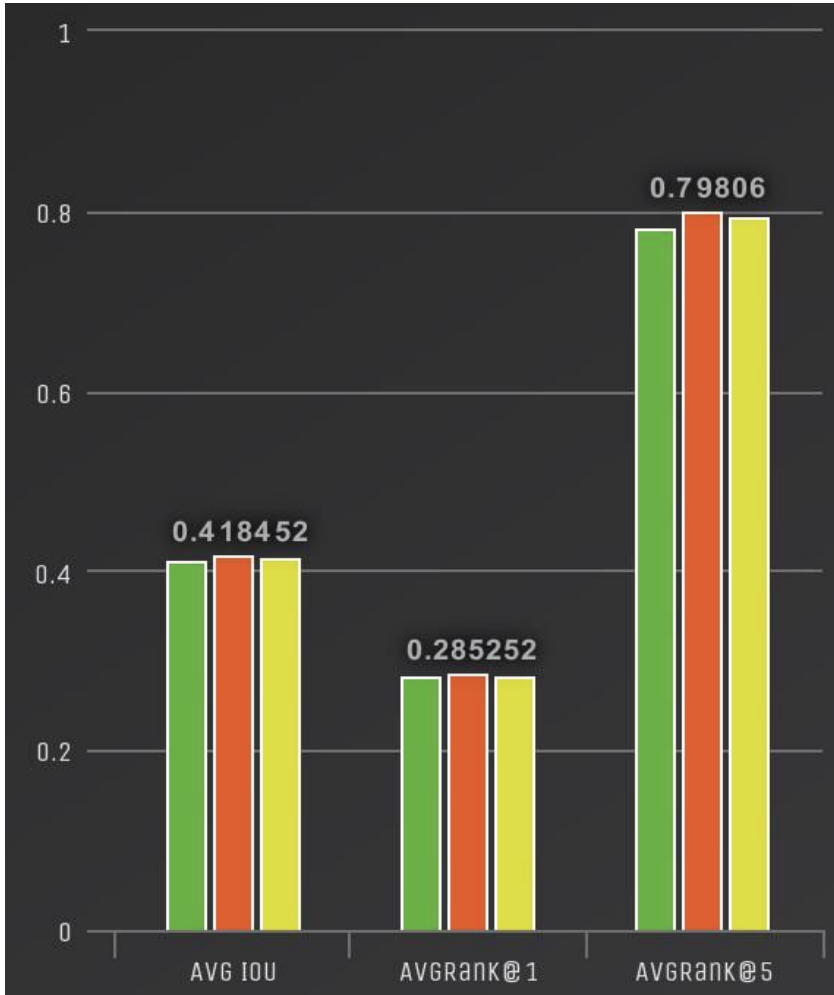
We first see people.

local+global+tef

local+global

Visual Features

Carnegie Mellon



1) **ResNet Max** : ResNet-50 instead of VGGNet + Optical Flow

Motivation : Max pooling, Richer Features

2) **Visual Fusion** : ResNet-50 Max + VGGNet Mean + Optical Flow

Motivation : Richer Visual Features

Audio Features - SoundNet - Conv5 Pool

Carnegie Mellon

Man lifts hand up before playing instrument

Predictions

Ground Truth - [0, 0] - 0-5 seconds

RGB + Flow - [1, 1] - 5-10 seconds

SoundNet - [0, 0] - 0-5 seconds

Audio Features

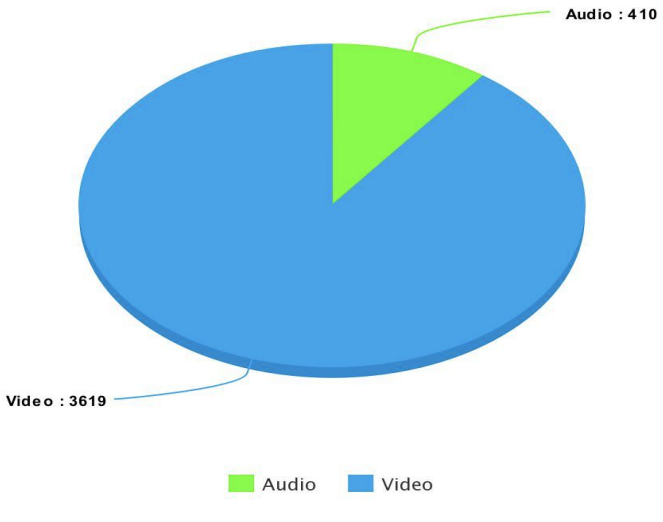
Old man sings in front of family

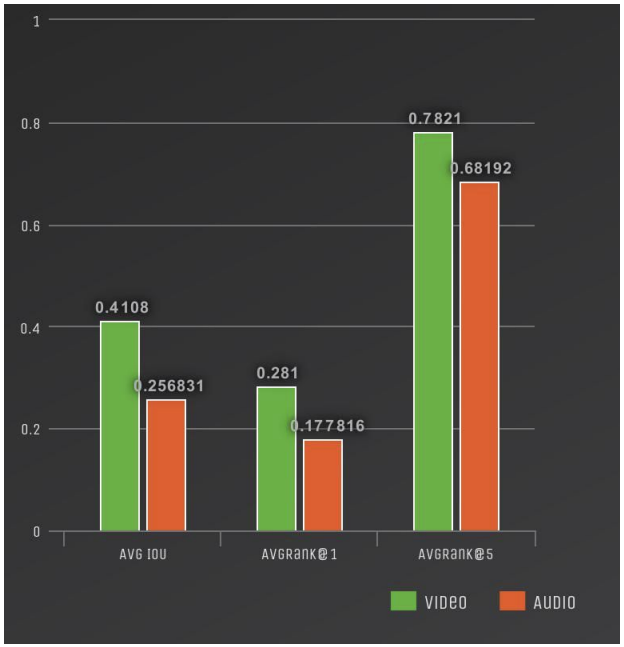
The shadow of the camera man can be seen

Audio vs Visual Features

Carnegie Mellon

WHY DID AUDIO FAIL ???

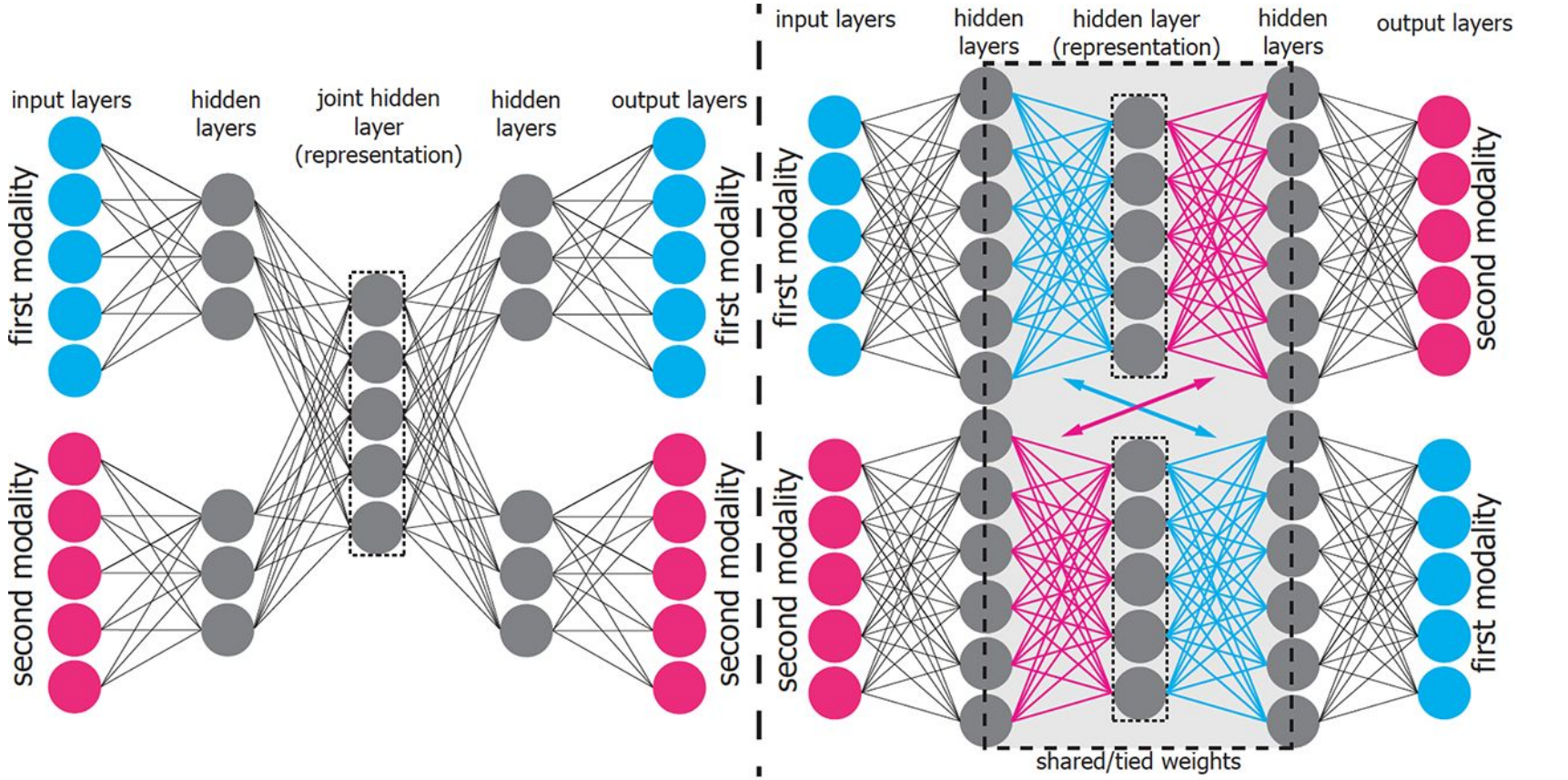




- Very few queries contain sound words
- Most queries deal with camera, spatial and temporal words
- No audio distinction between moments
- "Camera focuses on sign saying 'zombie disco disaster'"

Multimodal Fusion Technique - BiDNN

Carnegie Mellon

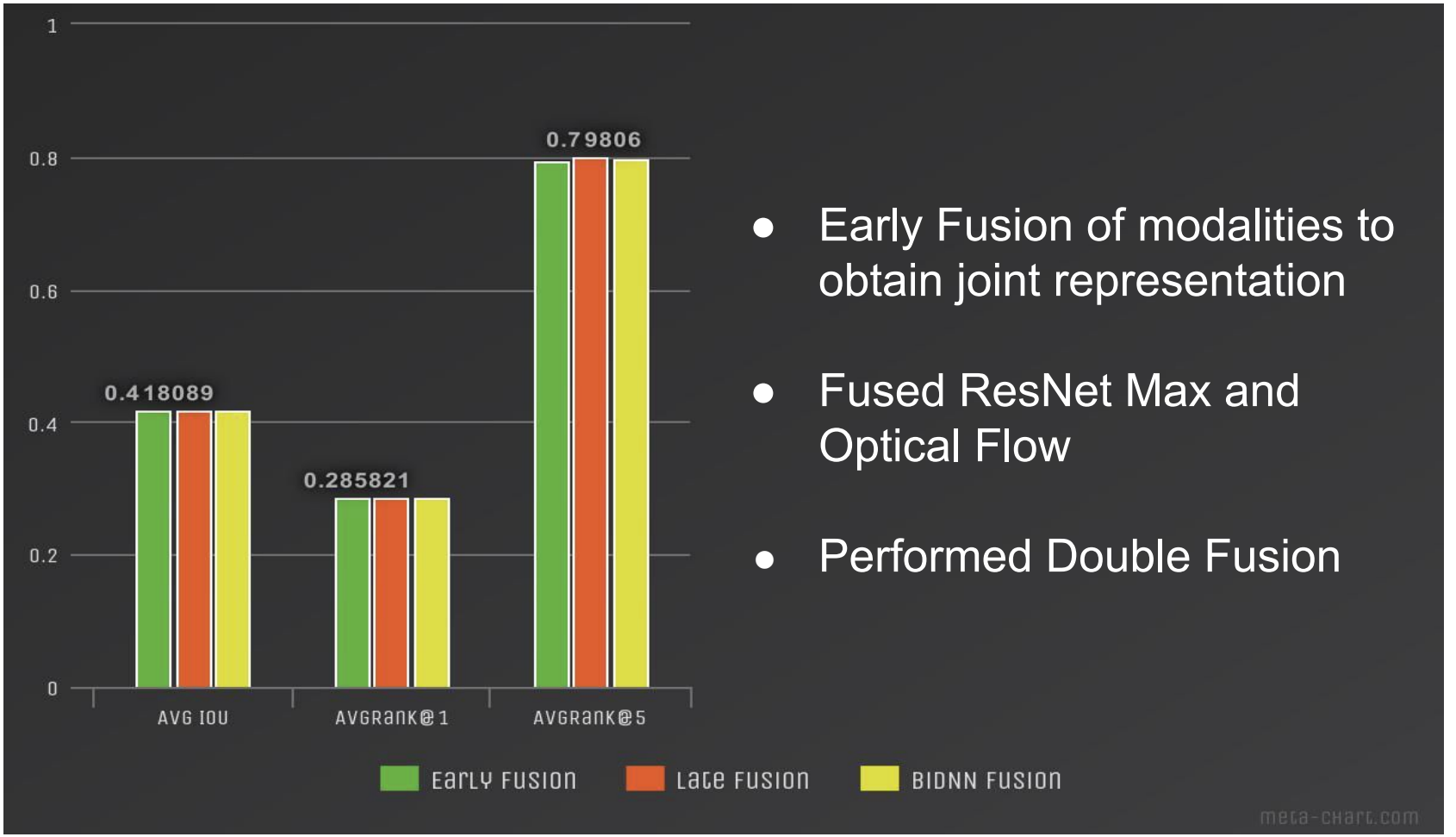


a) classical multimodal autoencoder

b) bidirectional (symmetrical) deep neural network(s)

BiDNN Results

Carnegie Mellon

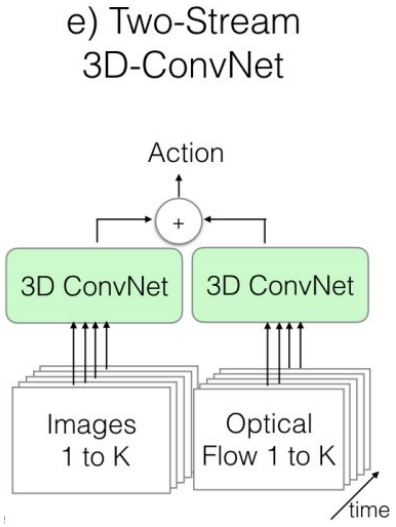


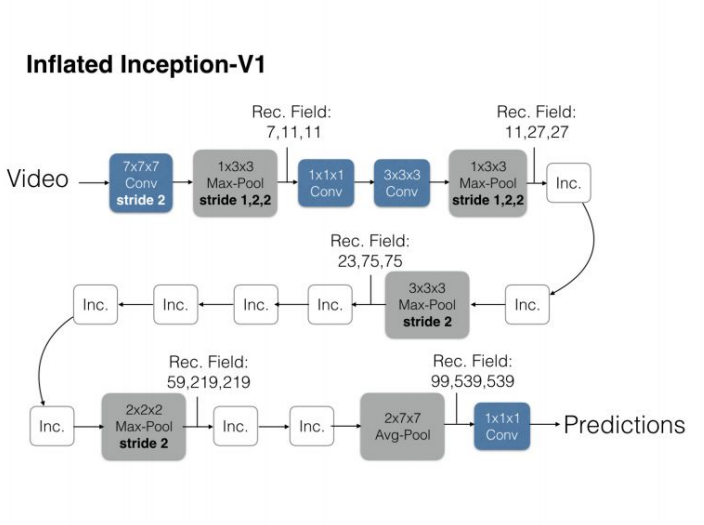
- Early Fusion of modalities to obtain joint representation
- Fused ResNet Max and Optical Flow
- Performed Double Fusion

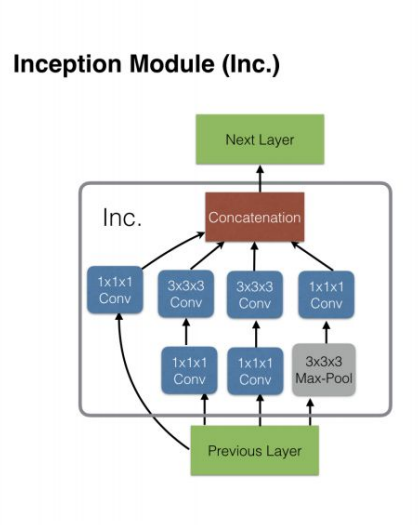
I3D - Two-Stream Inflated 3D Convnet

Carnegie Mellon

- Trained on Kinetics Dataset
- 400 human action classes, 400 clips per class







400 human activity classes - tossing coin, yawning, blowing out candles, etc

Activity Class: **Climbing Ladder**

Sample Queries:

- Small boy tries to climb ladder
- Little boy climbs green ladder
- Boy holds onto ladder
- A child in the playground holds the blue green ladder

I3D Results

Carnegie Mellon

Model	Avg IOU	AvgRank@1	AvgRank@5
I3D (Sent2Vec)	**0.39025**	**0.35978**	**0.94576**
I3D Feature (1024d)	**0.390741**	**0.361905**	**0.944974**
I3D Prediction (400d)	**0.3806**	**0.352381**	**0.939153**
Baseline (6 moments)	0.389771	0.359259	0.945767

Image Captioning - Future Work (Ongoing)

Carnegie Mellon

Activity recognition drawback:

-> class names are short phrases
-> human activities

Image captioning: Longer descriptions + generate any description

- A little boy climbs a small ladder to climb up the slide.
- The blue green ladder leads up the slide
- A small child in the playground climbs up the ladder to the slide

Match with query and calculate best match score

IMPORTANT TAKEAWAYS

Carnegie Mellon

1. Understanding the dataset and exploring dataset
2. Establishing baseline is extremely important
3. Beating baseline is not the sole objective
4. Exploring different visual and audio features
5. Multimodal fusion techniques - beyond concat and late fusion
6. Concepts like object identification, activity recognition, image captioning to learn models better.
7. Research and Networking
8. Time and resource management
9. Team dynamics

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.

- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, Andrew Zisserman. The Kinetics Human Action Video Dataset, arXiv:1705.06950, May 2017

- Vedran Vukotić, Christian Raymond, Guillaume Gravier. Multimodal and Crossmodal Representation Learning from Textual and Visual Features with Bidirectional Deep Neural Networks for Video Hyperlinking. ACM Multimedia 2016 Workshop: (iV&L-MM'16), Oct 2016, Amsterdam, Netherlands.

- Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. Show and tell: a neural image caption generator. In *Proc. International Conference on Machine Learning* http://arxiv.org/abs/1502.03044 (2014).

Q&A

Carnegie Mellon