INF@AVS 2018: Learning discrete and continuous representations for cross-modal retrieval

Po-Yao(Bernie) Huang, Junwei Liang, Vaibhav,
Xiaojun Chang and Alexander Hauptmann

Carnegie Mellon University, Monash University

Outline

- Introduction
- Discrete semantic representations for cross-modal retrieval
  - Conventional concept-bank approach
- Continuous representations for cross-modal retrieval
- Results and Visualization
  - 2016 results (http://vid-gpu7.inf.cs.cmu.edu:2016)
    - 12.6 mIAP v.s. 2017 AVS winner 10.2 mIAP (+ 23.5 %)
  - 2018 results (http://vid-gpu7.inf.cs.cmu.edu:2018)
    - 2nd place, 8.7 mIAP
- Discussion: What does/doesn't the model learn?
- Conclusion and future work

Visualization

http://vid-gpu7.inf.cs.cmu.edu:2016

http://vid-gpu7.inf.cs.cmu.edu:2018

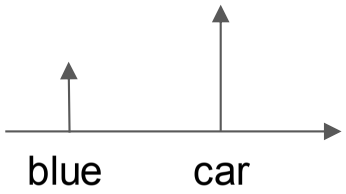
1507 => a choir or orchestra and conductor performing on stage

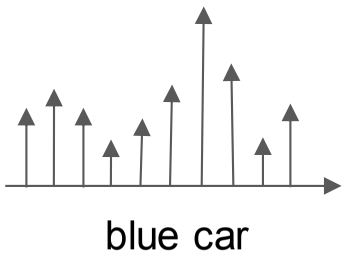
Sun 11 November 2018
By CMU

result_file:/home/poyaoh/scan_1000_t2i.txt

Introduction

- AVS as a cross-modal (text to video) retrieval problem
  - Vectorize representations for text queries and videos
    - $t_i = \text{encoder}_{text}(\text{query}_i)$, $v_j = \text{encoder}_{video}(\text{video}_j)$
  - Cross-modal retrieval based on distance between $\boldsymbol{t}, \boldsymbol{v}$.
    - $R(\boldsymbol{s}|q_i)$, $s_j = \text{dist}(v_j, t_i)$
- Two types of the joint embedding space $\boldsymbol{t}, \boldsymbol{v} \in R^N$
  - Discrete embeddings (Conventional approach with concept-bank)
    - Each dimension has a specific semantic meaning
  - Continuous embeddings
    - Each dimension doesn't have a specific meaning

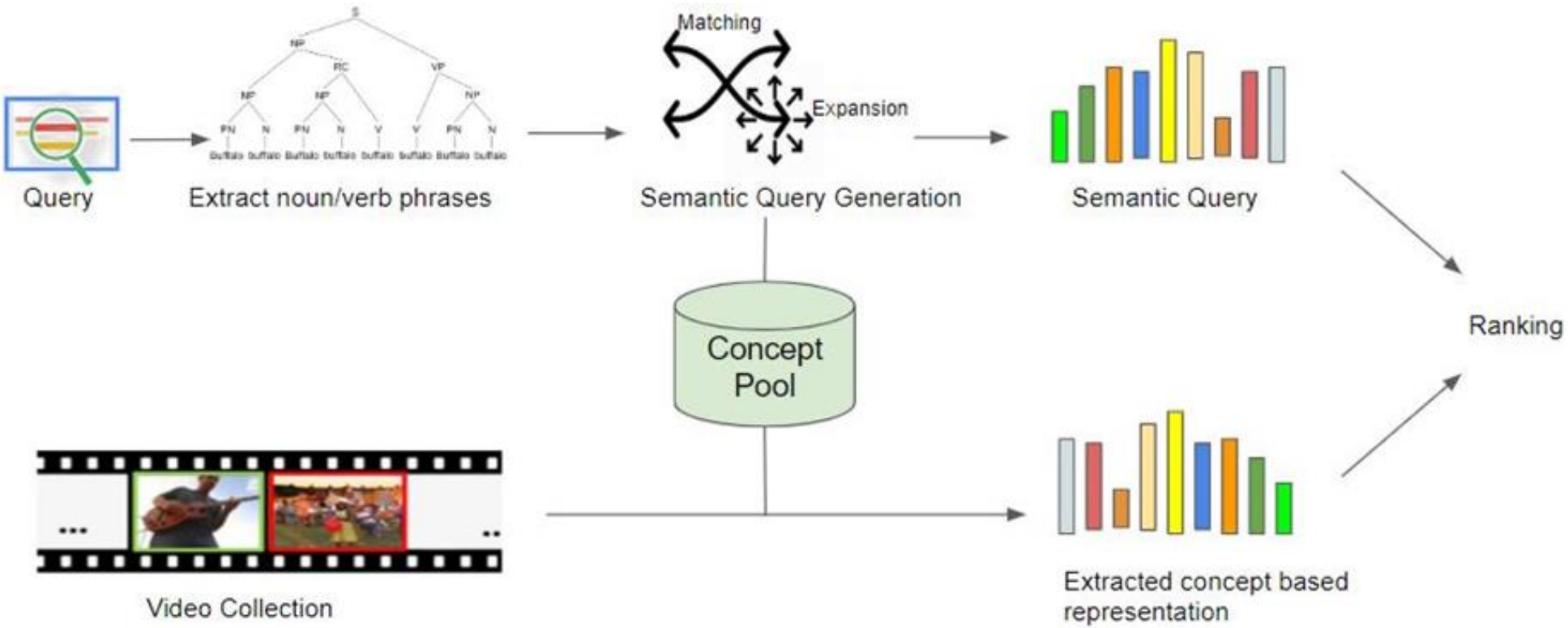




Introduction

- Discrete joint-embedding space: N: >10,000
  - Learnt from external (classification) dataset $\{(label, image/video)_i\}$
  - Pros: More interpretable. Easy to debug/re-rank
  - Cons: Less representation power, hard to generalize, curse of dimensionality (when N is large)
- Continuous joint-embeddings space: N: 500~1000
  - Learnt from external (retrieval/captioning) datasets with pairwise samples $\{(text, image/video)_i\}$
  - Pros: Usually more powerful, SOTA in multiple datasets
  - Cons: Not-interpretable, hard to control/debug
- AVS
  - Directly perform inference with the models pre-trained on external datasets to generate $\boldsymbol{t}$, $\boldsymbol{v}$
  - Output the ranking based on euclidean/cosine similarity scores

Pipeline for retrieval using discrete semantics



Two sub-problems when using discrete semantics

- Concept Extraction
  - Extract concepts from videos using pre-trained detectors
  - This can be done offline

- Semantic Query Generation (SQG)
  - Converting a text query to a concept vector
  - Given a new query, needs to be done online

Concept Extraction

- Datasets used for training concept detectors

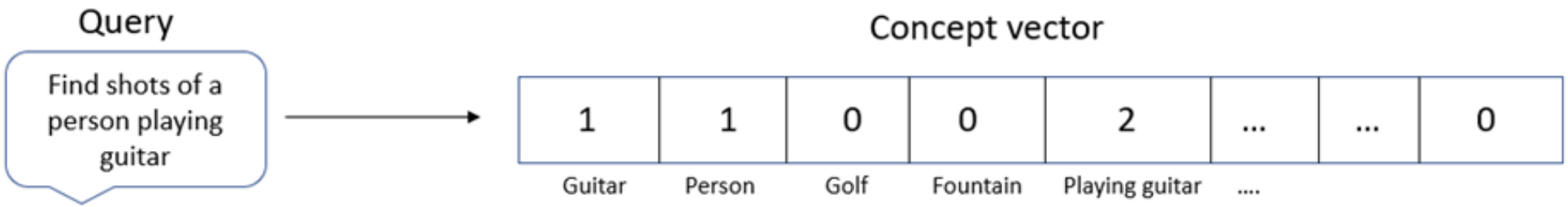
YFCC	609 concepts
ImageNet Shuffle	12703 concepts
UCF101	101 concepts
Kinetics	400 concepts
Place	365 concepts
Google Sports	478 concepts
FCVID	239 concepts
SIN	346 concepts
Moments	339 concepts

A total of 15,580 concepts in our **concept pool.**

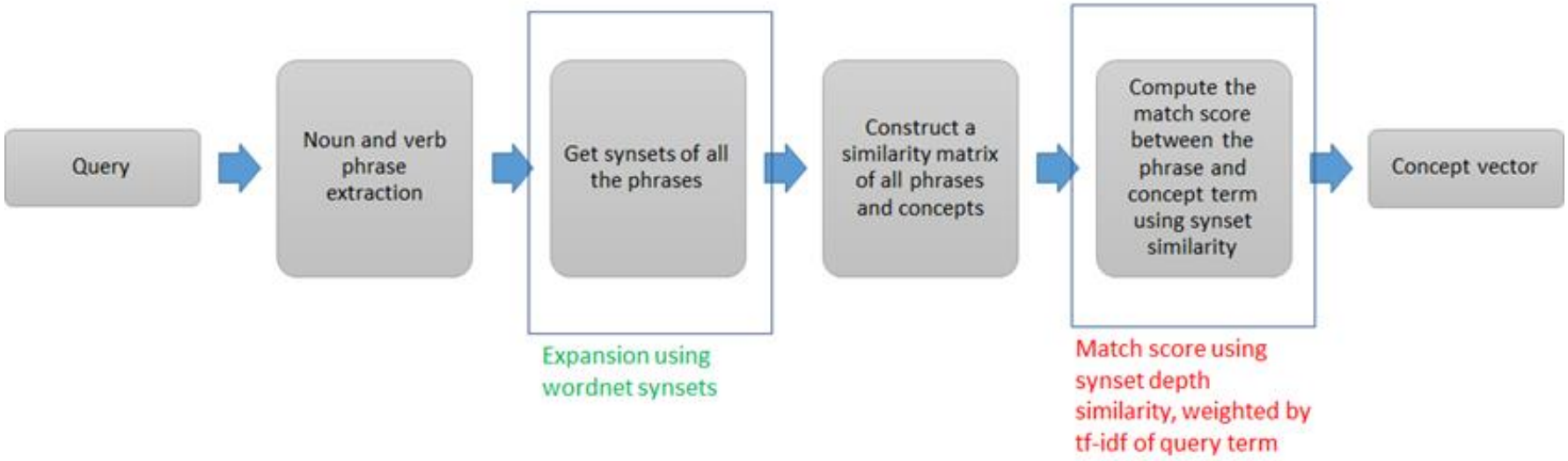
- Use these detectors offline to extract concepts from all the videos

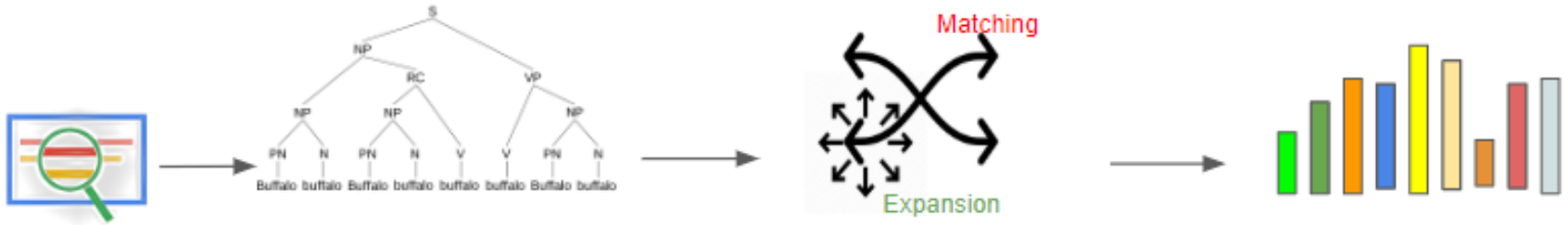
SQG Baseline: Exact Match

We convert a text query to a concept vector using **exact match** between the terms in query and concepts in the concept pool.



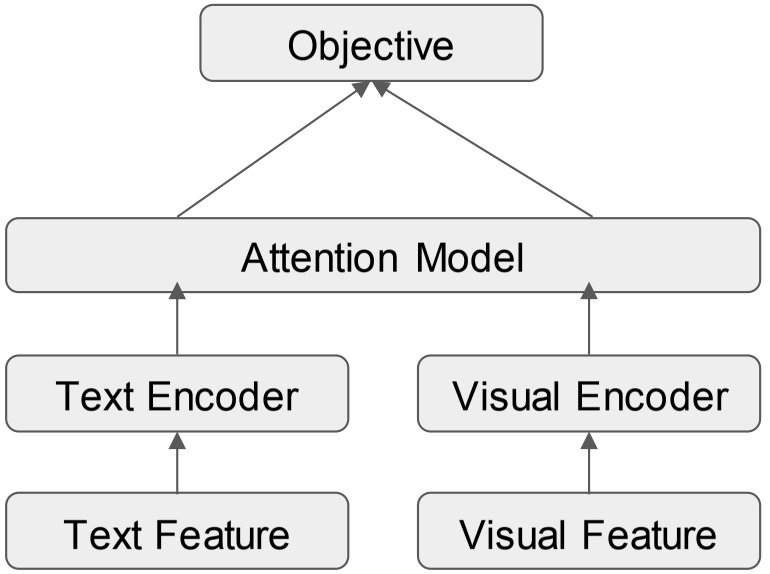
SQG: Synset Approach





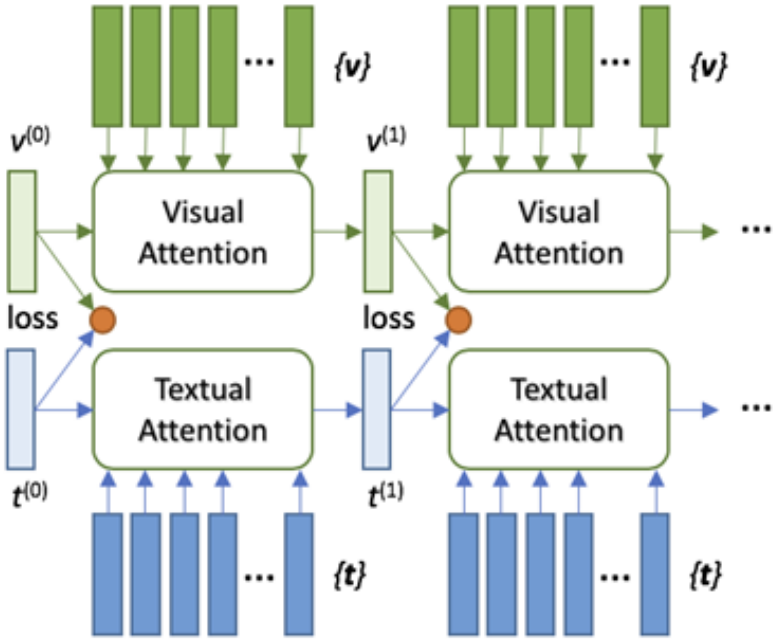
Models learning continuous embeddings

- Features and Encoders
  - Text encoder: GRU/LSTM
    - W2V: randomly initialized. Vocabulary: {Flickr30K ∪ MSCOCO ∪ MSR-VTT)
  - Visual encoder: A simple linear layer
    - Mean pooled frame-level regional features
      - Last Conv of ResNet 101
      - Last Conv of Faster RCNN (ResNet 101)
- Attention Model:
  - Intra-modal attention
  - Inter-modal attention
- Objective:
  - Pairwise max-margin loss
  - Hard negative mining

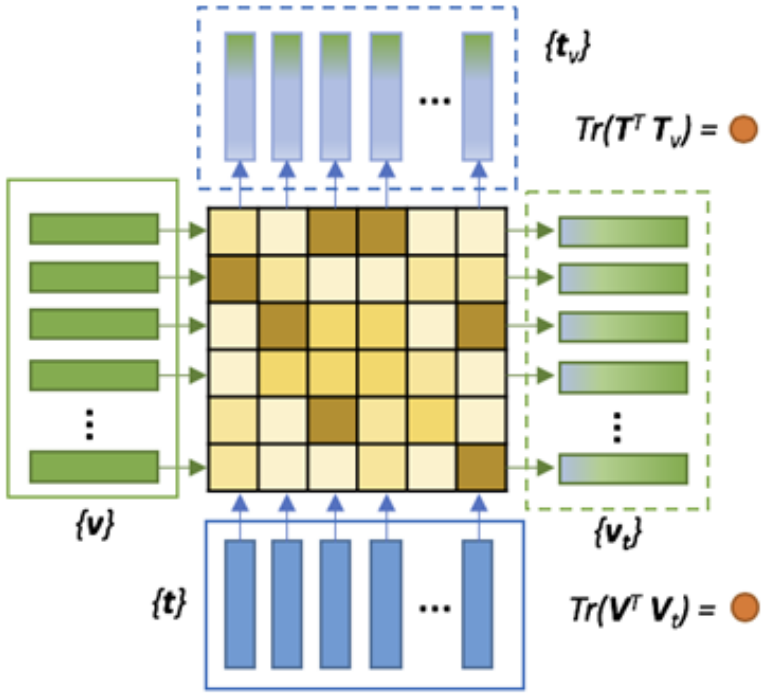


Models learning continuous embeddings

Intra-modal attention (DAN: Dual Attention Network)



Inter-modal attention (CAN: Cross Attention Network)

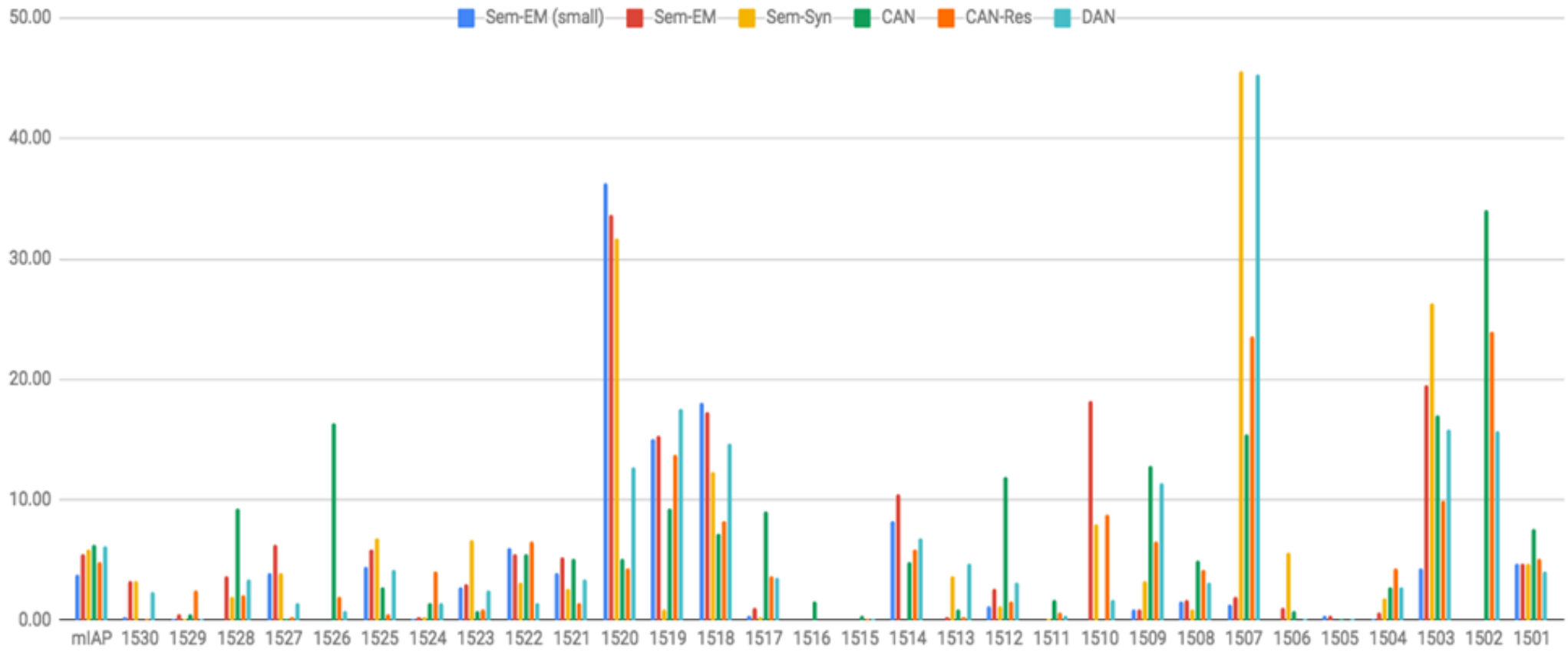


- Complexity at the inference phase: (M: # query, N: # data)
  - DAN (Intra-attention O(M))
  - CAN (Inter-attention O(MN))

Datasets and Experimental Settings

- Pre-trained dataset statistics
  - Flickr30K: 31,783 images, each with 5 text descriptions
  - MSCOCO: 123,287 images, each with 5 text descriptions (coco 2014)
  - MSR-VTT: 10,000 videos, each with 20 text descriptions
- Some hyperparameters
  - Embedding dim: 512, DAN # of hops: 2
  - Batch size 128, within-batch hardest negative mining
  - Adam optimizer with 0.001 learning rate, gamma 0.1 for 20 epochs, 50 epochs for training, 30 epochs for early stopping
- Features
  - 300-dim word embeddings, truncated at length 82.
  - 7x7x2048 for ResNet101, 36x2048 for faster-RCNN. Mean-pooled over frames in IACC.3.
- Fusion
  - Late fusion weights from Leave-one(model)-out. 11 models are fused.

Quantitative Results (IACC.3 2016)



Quantitative Results

- 1510: a sewing machine
- 1512: palm trees
- 1518: one or more people at train station platform
- 1520: any type of fountains outdoors
- 1526: a woman wearing glasses
- 1529: a person lightening a candle
- Fusion weights: (11 models)
  - Discrete: 0.53 (5 models)
  - Continuous: 0.47 (6 models)

Qeury	Sem-EM (small)	Sem-EM	Sem-Syn	CAN	CAN-Res	DAN	Fusion	
1501	4.77	4.76	4.75	7.66	5.11	4.05		-
1502	0.00	0.01	0.00	34.12	24.00	15.68		-
1503	4.33	19.51	26.38	17.01	9.93	15.92		-
1504	0.19	0.63	1.87	2.78	4.28	2.78		-
1505	0.40	0.37	0.00	0.07	0.05	0.09		-
1506	0.00	1.01	5.66	0.74	0.01	0.07		-
1507	1.37	2.01	45.64	15.44	23.64	45.35		-
1508	1.54	1.65	0.88	5.03	4.14	3.20		-
1509	0.89	0.85	3.28	12.86	6.49	11.45		-
1510	0.00	18.17	8.03	0.01	8.77	1.69		-
1511	0.04	0.05	0.18	1.75	0.69	0.38		-
1512	1.15	2.56	1.23	11.95	1.57	3.19		-
1513	0.06	0.23	3.73	0.91	0.26	4.67		-
1514	8.28	10.45	0.00	4.81	5.96	6.85		-
1515	0.03	0.04	0.05	0.35	0.16	0.10		-
1516	0.00	0.00	0.01	1.58	0.00	0.00		-
1517	0.44	1.07	0.22	9.00	3.71	3.56		-
1518	18.14	17.36	12.28	7.25	8.26	14.69	?	-
1519	15.04	15.27	0.85	9.27	13.75	17.58		-
1520	36.26	33.64	31.77	5.16	4.34	12.67	?	-
1521	3.91	5.21	2.58	5.12	1.40	3.39		-
1522	6.00	5.44	3.08	5.49	6.58	1.44		-
1523	2.71	2.96	6.63	0.72	0.88	2.48		-
1524	0.14	0.24	0.30	1.50	4.11	1.49		-
1525	4.43	5.92	6.82	2.69	0.49	4.13		-
1526	0.00	0.04	0.00	16.42	2.00	0.73		-
1527	3.92	6.29	3.90	0.15	0.23	1.40		-
1528	0.00	3.70	1.90	9.27	2.04	3.35		-
1529	0.09	0.53	0.12	0.46	2.44	0.08		-
1530	0.29	3.26	3.33	0.02	0.09	2.39		-
IAP	**3.81**	**5.44**	**5.85**	**6.32**	**4.85**	**6.16**		**12.59**

Qualitative results on AVS 2016 queries

1510 Find shots of a sewing machine

CAN: 0.01

SYN: 8.03 (sewing machine in the semantic pool)

1512 Find shots of palm trees

CAN: 11.95

SYN: 1.23 (palm trees: OOV)

1526 Find shots of a woman wearing glasses

CAN: 16.42 (understands “wearing glasses” and woman)

SYN: 1.23 (disambiguation of matching/ SQG fails)

1529 Find shots of a person lighting a candle

CAN: 0.46 (

SYN: 0.53

1507 Find shots of a choir or orchestra and conductor performing on stage

CAN: 11.95

SYN: 45.24

1518 one or more people at train station platform

CAN: 7.25 ??

SYN: 45.24

Qualitative results on AVS 2018 queries

Find shots of people waving flags outdoors

CAN:

SYN:

Find shots of one or more people hiking

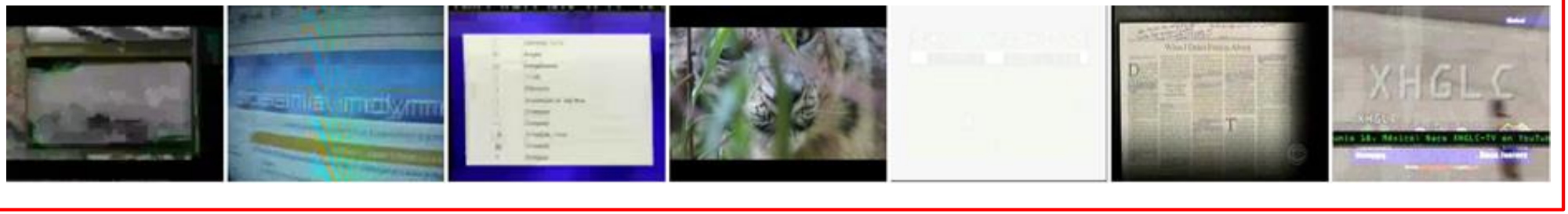
CAN:

SYN:

Find shots of a projection screen

CAN:

EM:



Find shots of a projection screen

SYN:

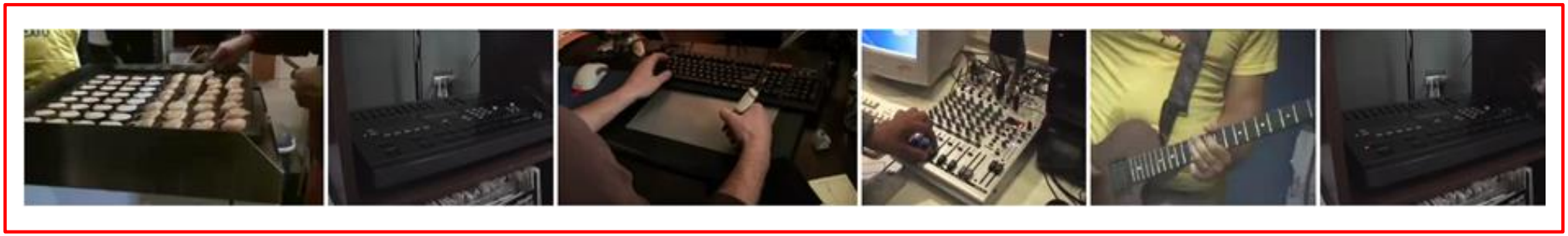
EM:

Find shots of a person sitting on a wheelchair

CAN:

SYN:

Find shots of a person playing keyboard and singing indoors



Discussion: What does/doesn't the model learn?

- Q: Does discrete semantics generalize for cross-modal retrieval?
- A: Probably NO without domain adaptation.
- Experiment:
  - Using the discrete representation (semantic concept bank) for text-to-image retrieval on Flickr30K
  - Results:

Model	R@1	R@5	R@10
Discrete semantics	6.1	17.7	22.4
CAN from coco (no training)	21.7	36.5	55.2
Published SOTA (CAN)	45.8	74.4	83.0
Ours (to be published)	53.3	80.0	85.4

Discussion: What does/doesn't the model learn?

prior

POS	Counts	Docs	All Docs
CD	39,611	33,873	158,915
VB	284,669	145,347	158,915
NN	650,207	158,904	158,915
JJ	195,647	107,660	158,915

- Q: What does /doesn't the continuous model learn?
- A: It cares nouns >>> adjs >> verbs > order > count.
  Syntactics, counting, preprop… in the text query should but does NOT matter...
- Experiment: (A simplified Intra-modal attention model)
  - Dropping/ shuffling text queries and compare how much does the performance drop

	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
baseline					44.68	73.30	79.26	57.5
random	42.04	72.92	79.06	55.72	42.76	72.48	78.21	56.05
no count	43.82	72.76	79.20	56.61	43.62	72.76	78.91	56.54
no verb	42.91	71.02	76.86	55.62	42.36	70.00	76.18	55.03
no noun	1.22	4.26	6.31	3.42	0.93	3.06	4.64	2.78
no adj	40.42	69.58	76.94	53.64	39.56	68.74	75.96	52.76

Table 3: Importance of different POS tags

Conclusion & future work

- We explored models learning two types of joint-embedding space for text to video retrieval for AVS
- Discrete semantics are good at finding specific (dominating) concept but are sensitive to OOV. They highly depend on the domain and are relatively hard to generalize to other datasets.
- Models with continuous embeddings are good at capturing latent/ compositional concepts and are complementary to the discrete models.
- Current SOTA cross-modal retrieval models learns mainly aligning nouns (objs) and adjs but care less about syntactics, counting.
- Combining the pros of two types of the model is our next step.