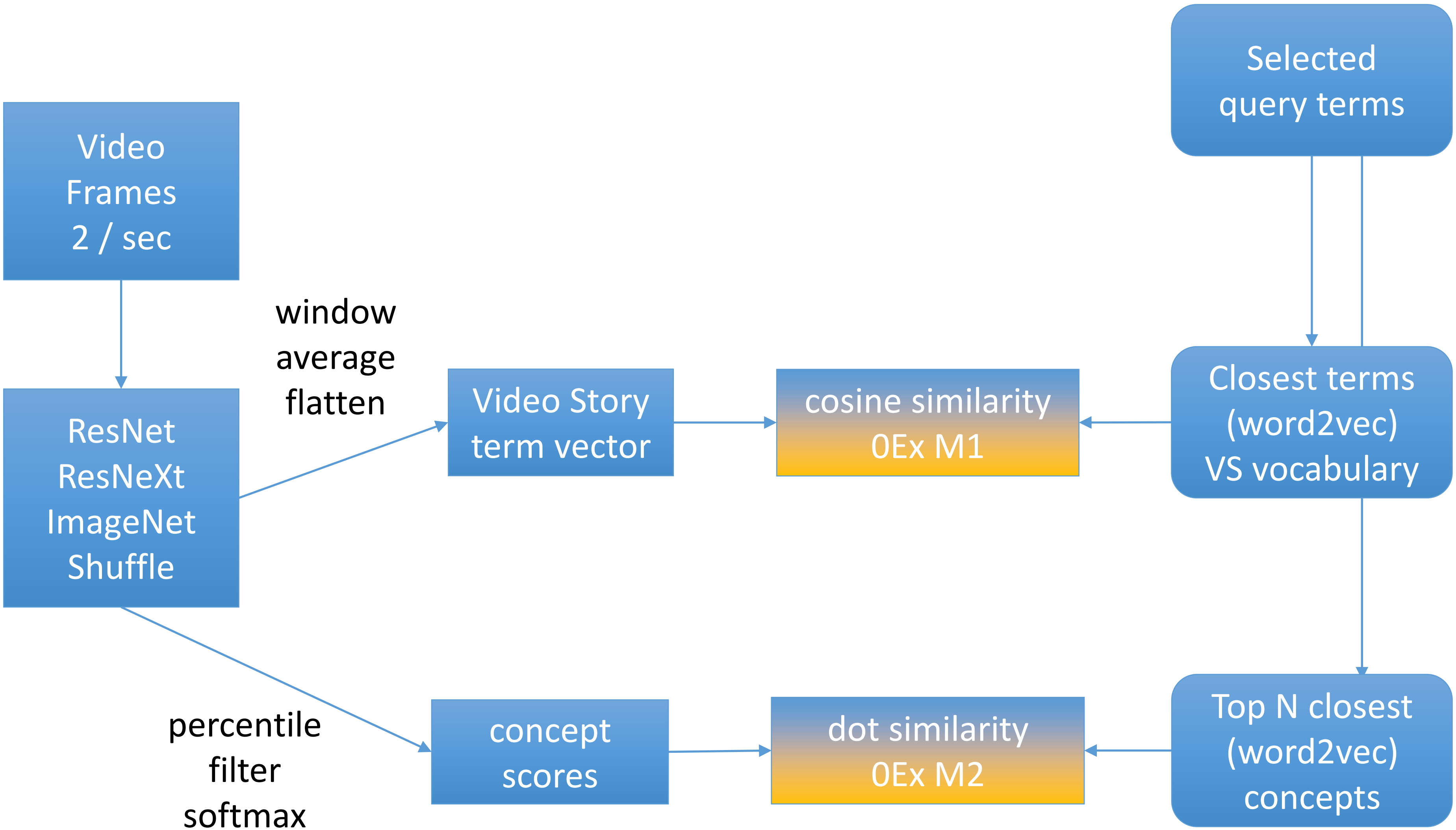
Query Understanding is Key for Zero-Example Video Search

Dennis Koelma and Cees Snoek

University of Amsterdam

The Netherlands

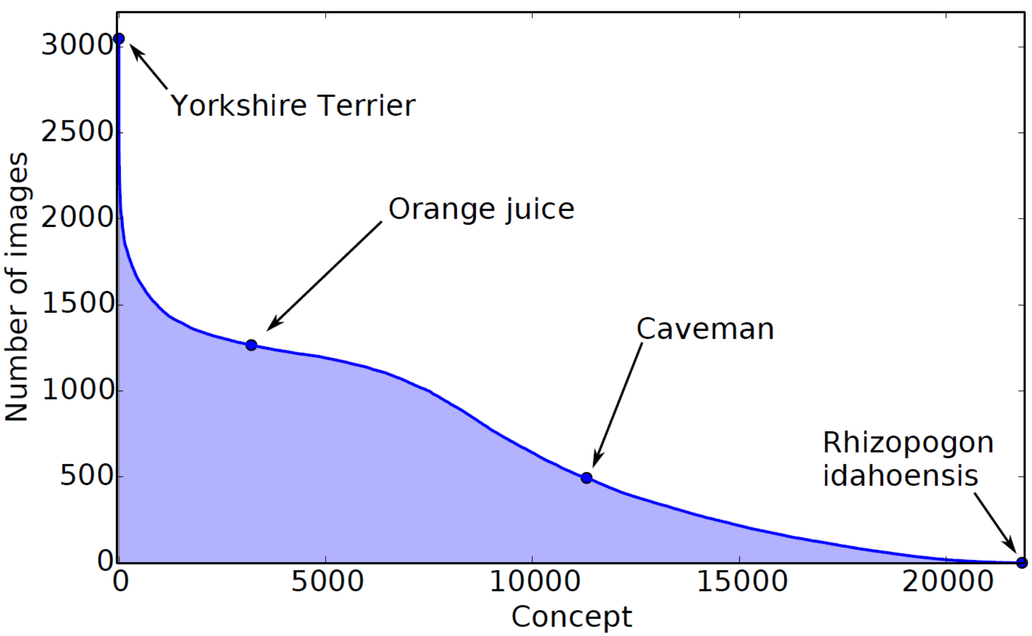
Pipeline


Video Frames 2 / sec
ResNet ResNeXt ImageNet Shuffle
window average flatten
Video Story term vector
cosine similarity 0Ex M1
Selected query terms
Closest terms (word2vec) VS vocabulary
percentile filter softmax
concept scores
dot similarity 0Ex M2
Top N closest (word2vec) concepts

22k ImageNet classes

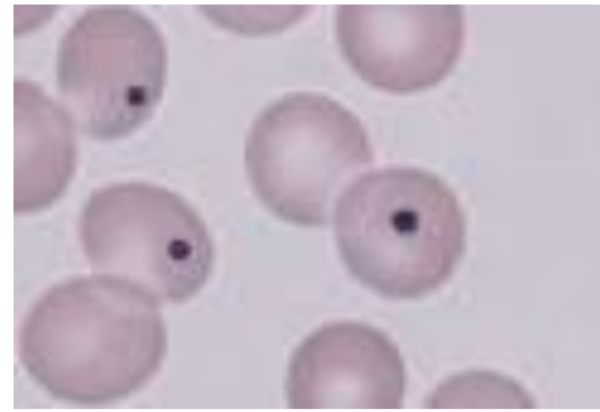
- Use as many classes as possible
- Find a balance between level of abstraction of classes and number of images in a class

Example imbalance



296 classes with 1 image

Irrelevant classes

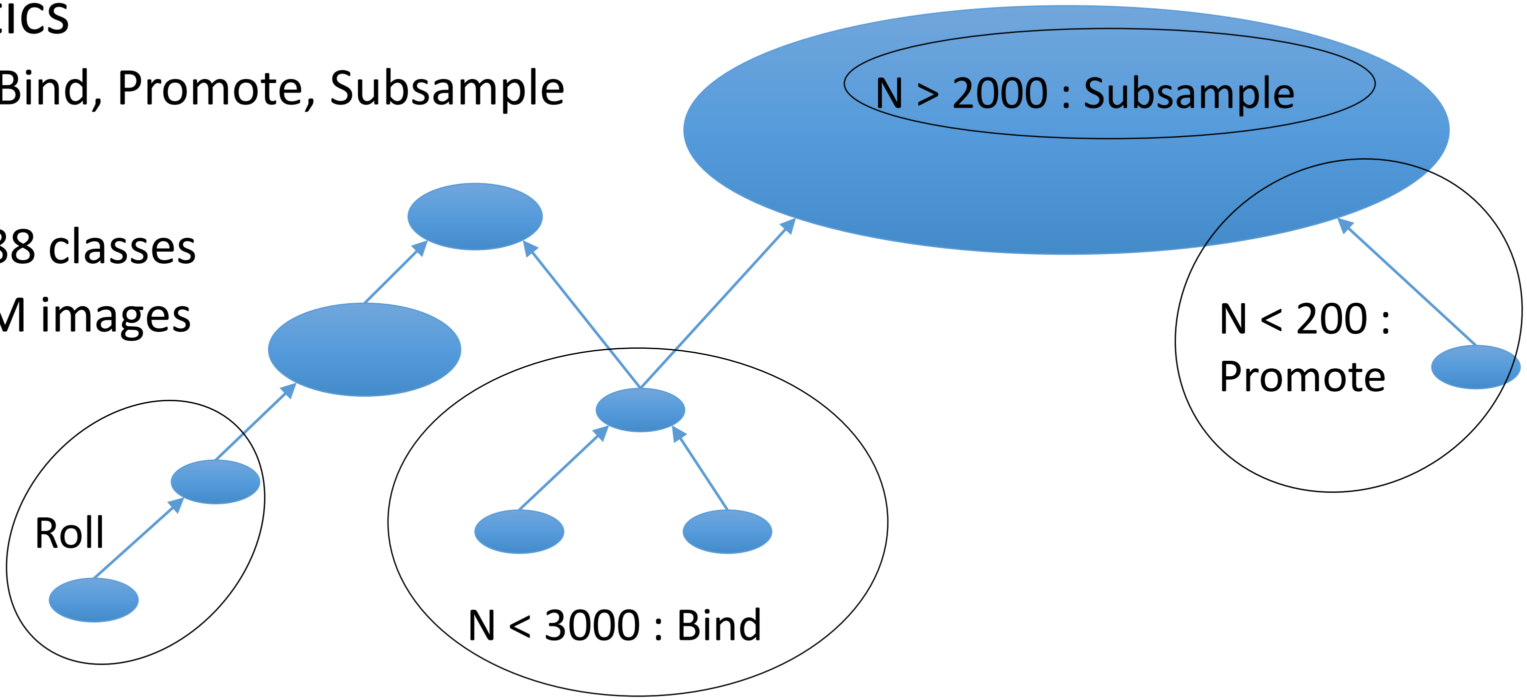


Siderocyte

Gametophyte

CNN training on selection out of 22k ImageNet classes

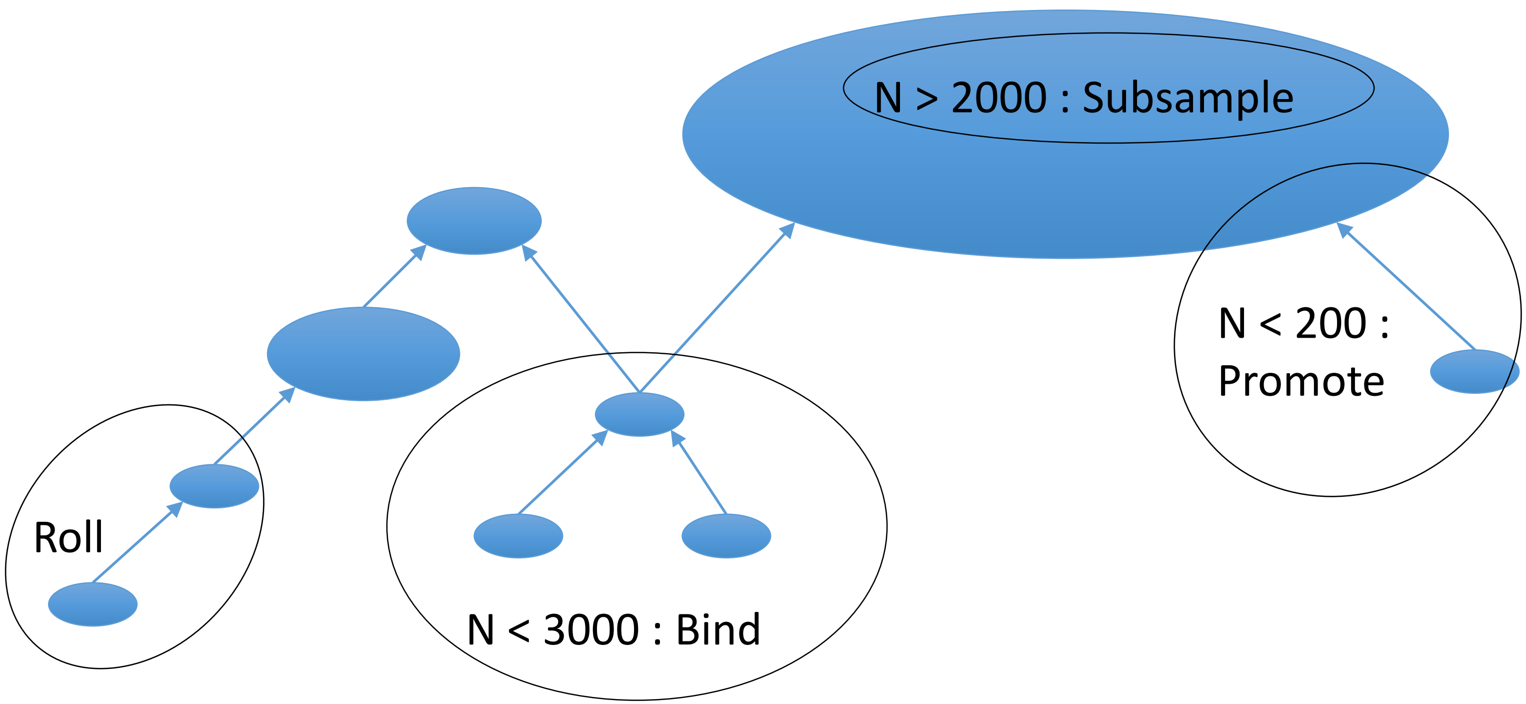
- Idea
  - Increase level of abstraction of classes
  - Incorporate classes with less than 200 samples
- Heuristics
  - Roll, Bind, Promote, Subsample
- Result
  - 12,988 classes
  - 13.6M images



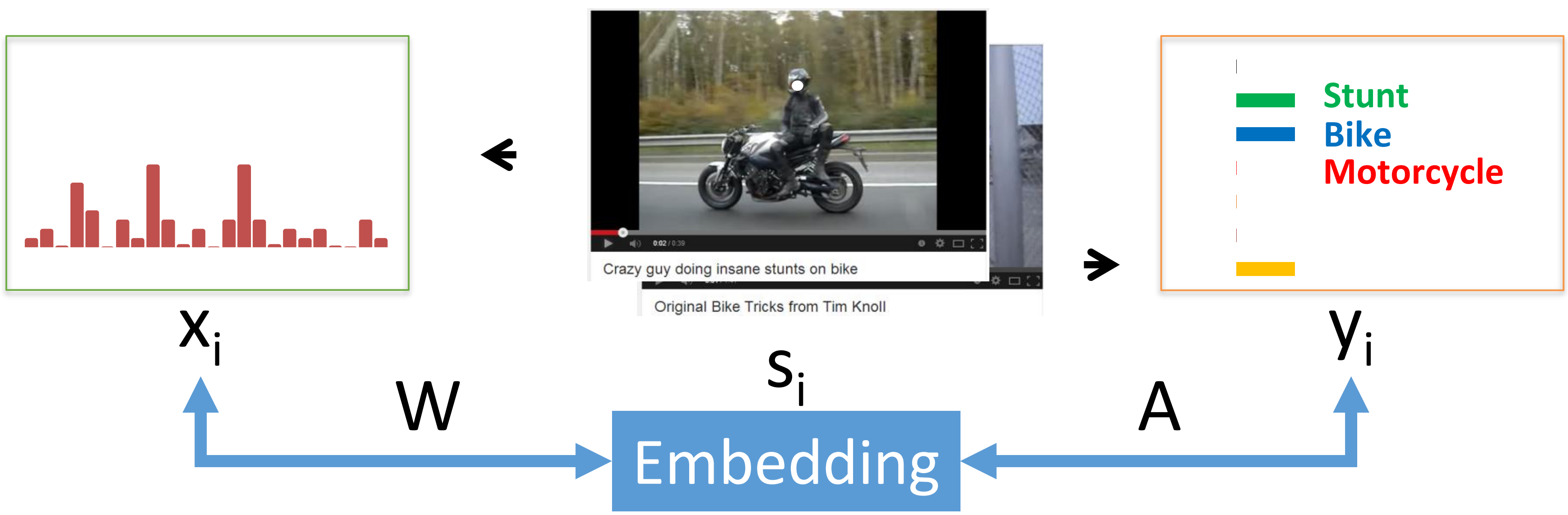
The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection,
Pascal Mettes and Dennis Koelma and Cees Snoek,
International Conference on Multimedia Retrieval, 2016

Concept Bank

- Two networks
  - ResNet
  - ResNeXt
- Three datasets (subsets of ImageNet)
  - Roll Bind (3000) Promote (200) Subsample, 13k classes, training: 1000 images/class
  - Roll Bind (7000) Promote (1250) Subsample, 4k classes, training: 1706 images/class
  - Top 4000 classes, Breadth-first search >1200 images, training: 1324 images/class



Video Story: Embed the story of a video



Joint optimization of W and A to preserve

Descriptiveness: preserve video descriptions : L(A,S)

Predictability: recognize terms from video content : L(S,W)

Videostory: A new multimedia embedding for few-example recognition and translation of events,
Amirhossein Habibian and Thomas Mensink and Cees Snoek,
Proceedings of the ACM International Conference on Multimedia, 2014

Video Story Training Sets

- VideoStory46k - www.mediamill.nl
  - 45826 videos from YouTube based on 2013 MED research set terms
- FCVID: Fudan Columbia Video Dataset
  - 87609 videos
- EventNet
  - 88542 videos
- Merged (VideoStory46k, FCVID, EventNet)

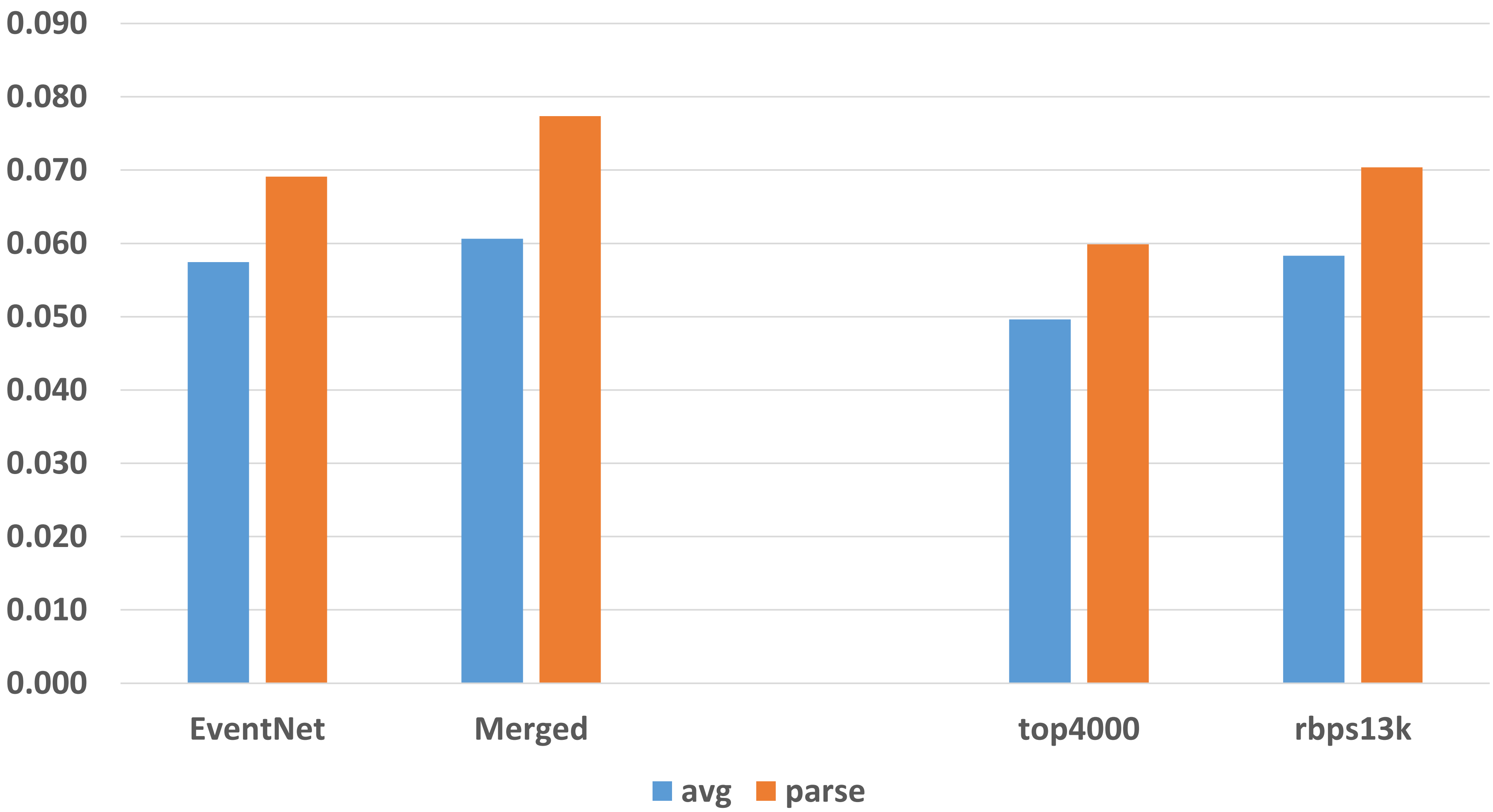
- Video Story dictionary: Terms that occur more than 10 times in the dataset
  - Merged : 6440 terms
- Using vocabulary of stemmed terms that occur more than 100 times in Wikipedia dump
  - With stemming: Respect the Video Story dictionary
  - 267.836 terms
- Use word2vec to expand them per video

Query Terms

- Experiments show it is important to select the right terms
  - Instead of just taking the average of the terms in word2vec space
- Part-of-Speech tagging
  - <noun1> , <verb> , <noun2>
  - <subject> , <predicate> , <remainder>
- Query Plan
  A. Use nouns, verbs, and adjectives in <subject>
     - unless it concerns a person (noun1 = “person”, ”man”, “woman”, “child”, …)
  B. Use nouns in <remainder>
     - unless it concerns a person or noun is a setting (“indoors”, “outdoors”, …)
  C. Use <predicate>
  D. Use all nouns in sentence
     - Unless noun is a person or a setting

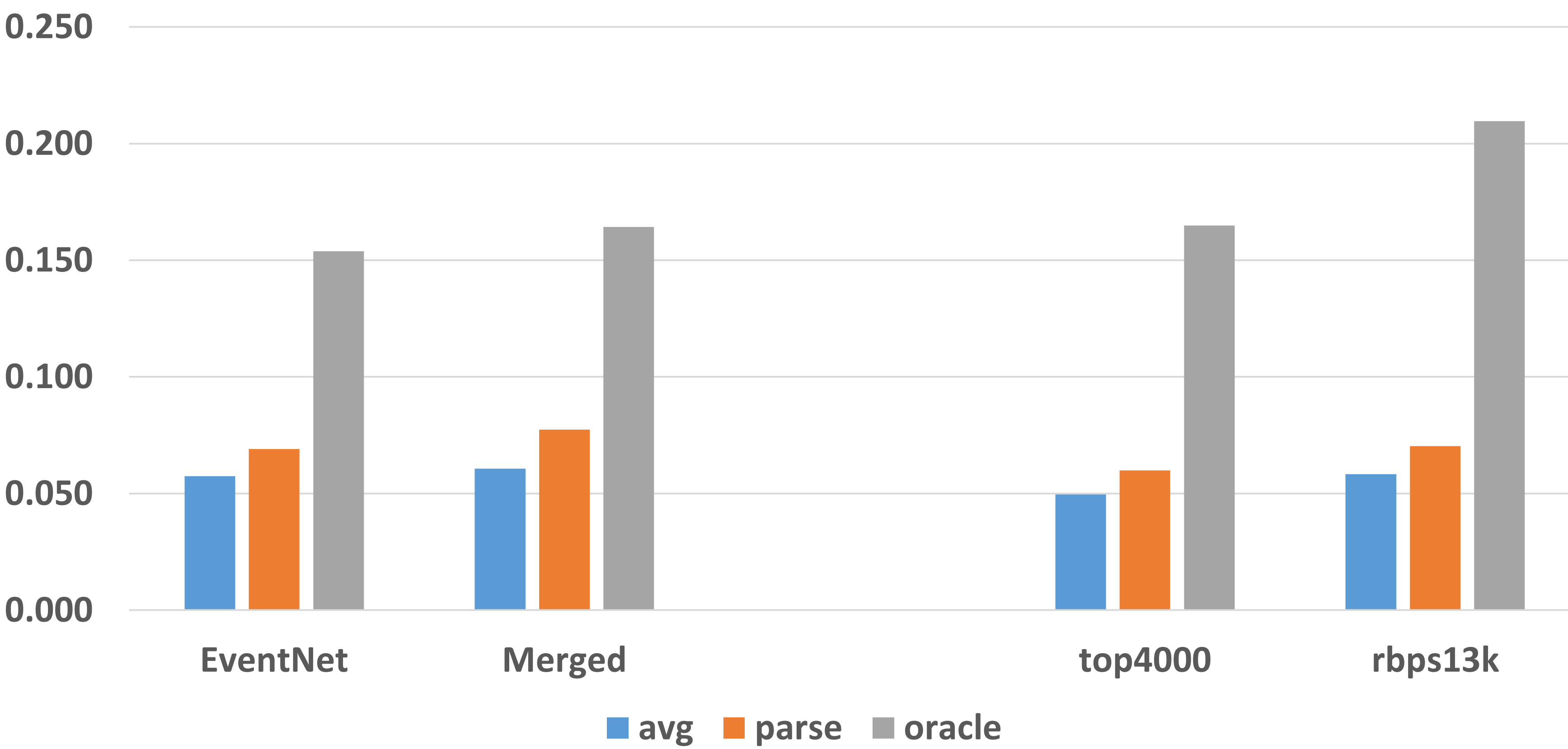
The Effect of Parsing on 2016 Topics

- MIAP using only ResNet feature



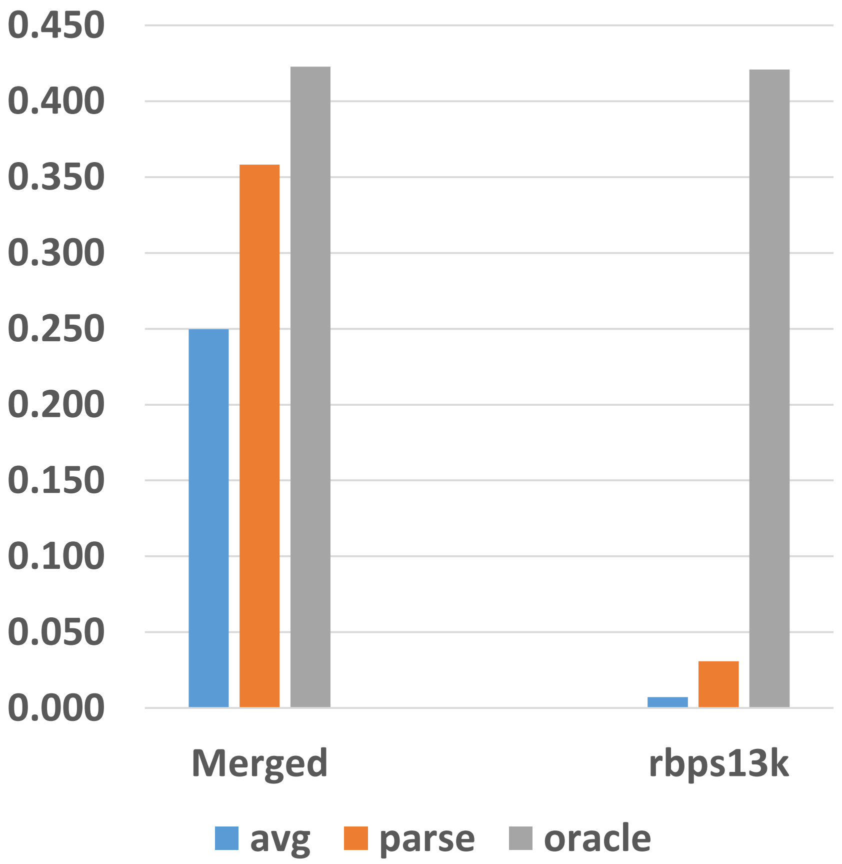
(Greedy) Oracle on 2016 Topics

- Fuse top (max 5) words/concepts with highest MIAP
- MIAP using only ResNet feature



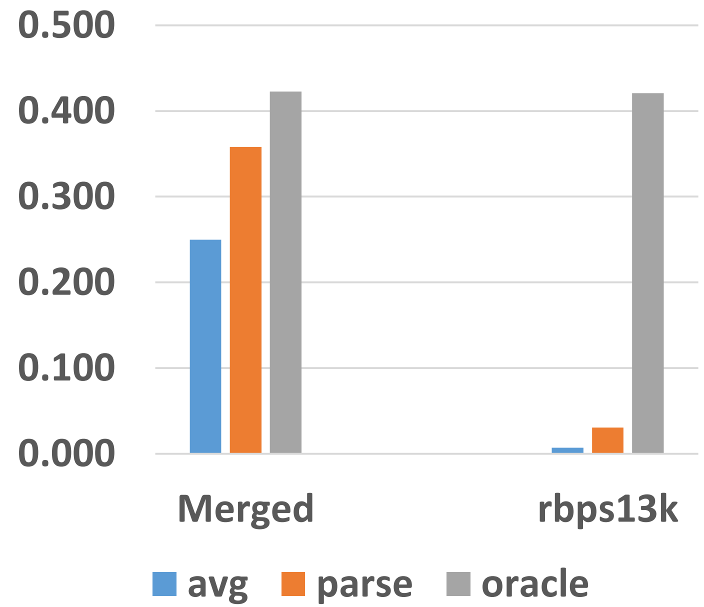
Query Examples : The Good

- A person playing drums indoors
- VideoStory terms avg :
  - person
  - plai
  - drum
  - indoor
- VideoStory terms parse :
  - drum
- VideoStory terms oracle :
  - beat
  - drum
  - snare
  - vibe
  - bng



Query Examples : The Ambiguous

- A person playing drums indoors
- Concepts top5 avg :
  - guitarist, guitar player
  - outdoor game
  - drum, drumfish
  - sitar player
  - brake drum, drum
- Concepts top5 parse :
  - drum, drumfish
  - brake drum, drum
  - barrel, drum
  - snare drum, snare, side drum
  - drum, membranophone, tympan

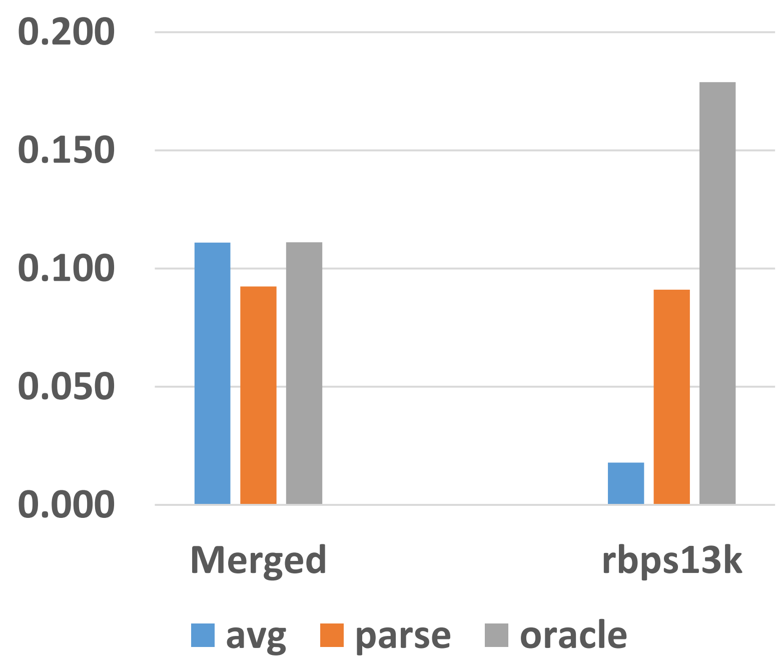


Oracle :

percussionist
cymbal
drummer
drum, membranophone, tympan
snare drum, snare, side drum

Query Examples : The Bad

- A person sitting down with a laptop visible
- VideoStory terms avg :
  - person
  - sit
  - laptop
- VideoStory terms parse :
  - laptop
- VideoStory terms oracle :
  - monitor
  - aspir
  - acer
  - alienwar
  - vaio
  - asus
  - laptop (rank 7)



Query Examples : The Difficult

- A person wearing a helmet
- Concept top5 parse :

  helmet (a protective headgear made of hard material to resist blows)

  helmet (armor plate that protects the head)

  pith hat, pith helmet, sun helmet, topee, topi

  batting helmet

  crash helmet
- Concept top5 oracle :

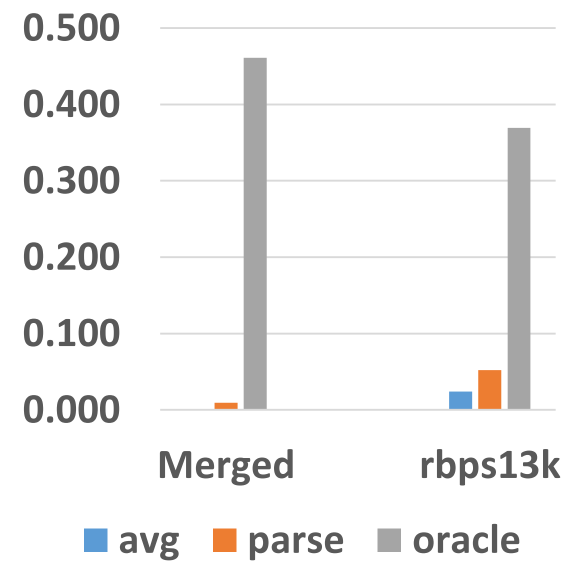
  hockey skate

  hockey stick

  ice hockey, hockey, hockey game

  field hockey, hockey

  rink, skating rink



Query Examples : The Impossible

- A crowd demonstrating in a city street at night
  - Parsing “fails”
  - Average wouldn’t have helped
- VS oracle :

  vega

  squar

  gang

  times

  occupi
- Concept oracle :

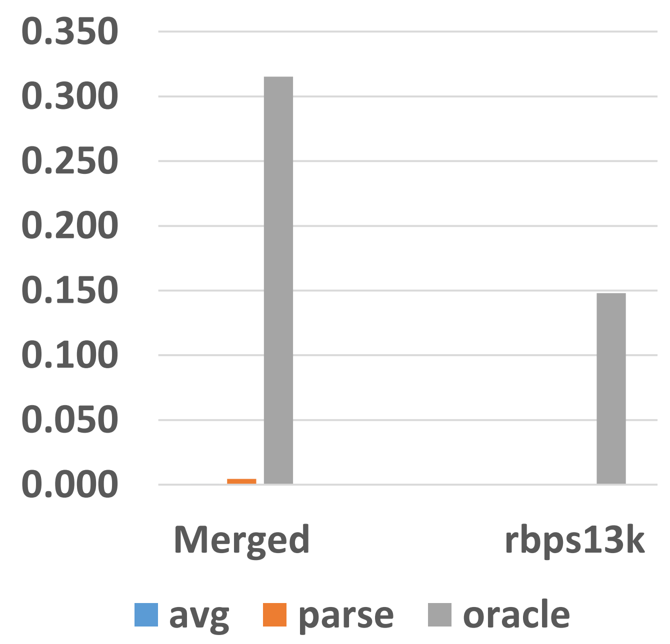
  vigil light, vigil candle

  motorcycle cop, motorcycle policeman, speed cop

  rider

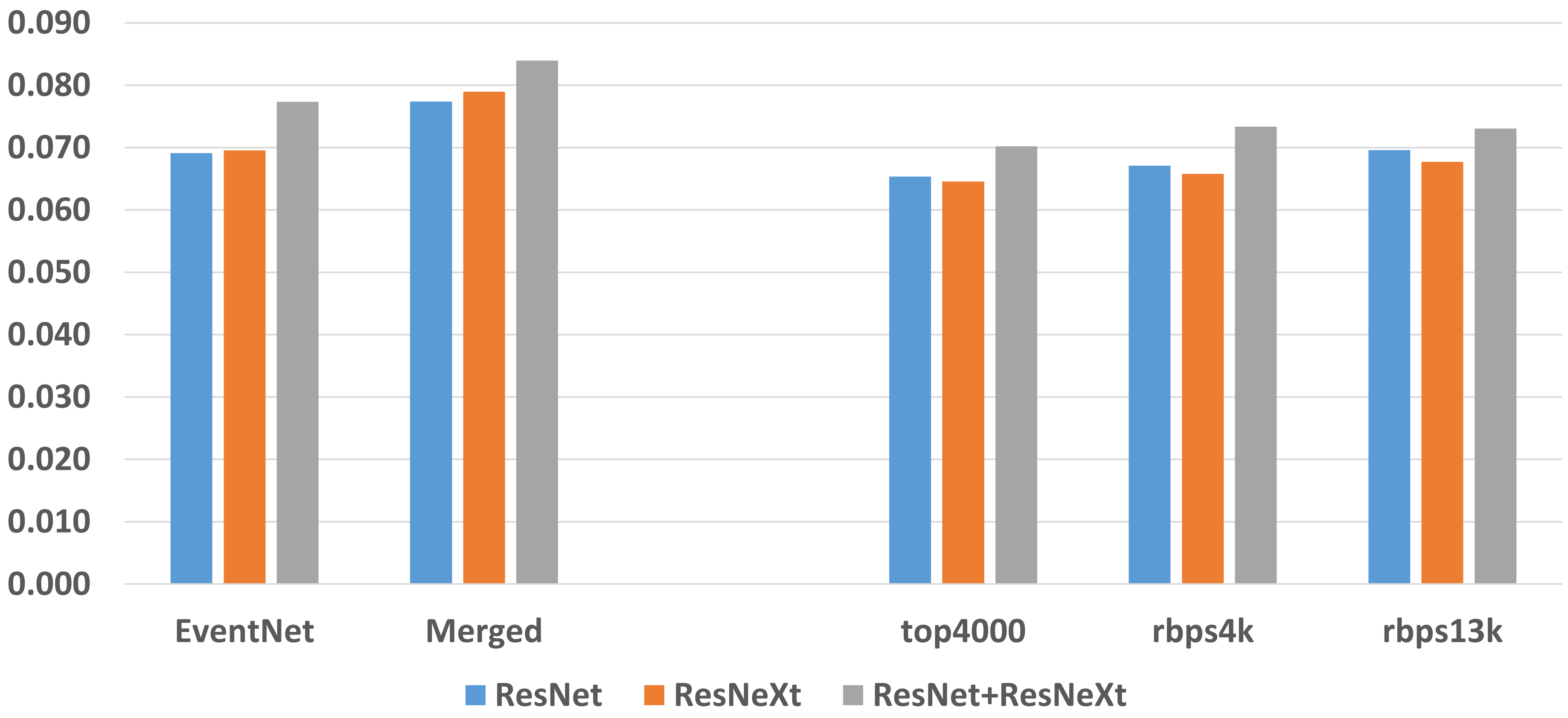
  minibike, motorbike

  freewheel



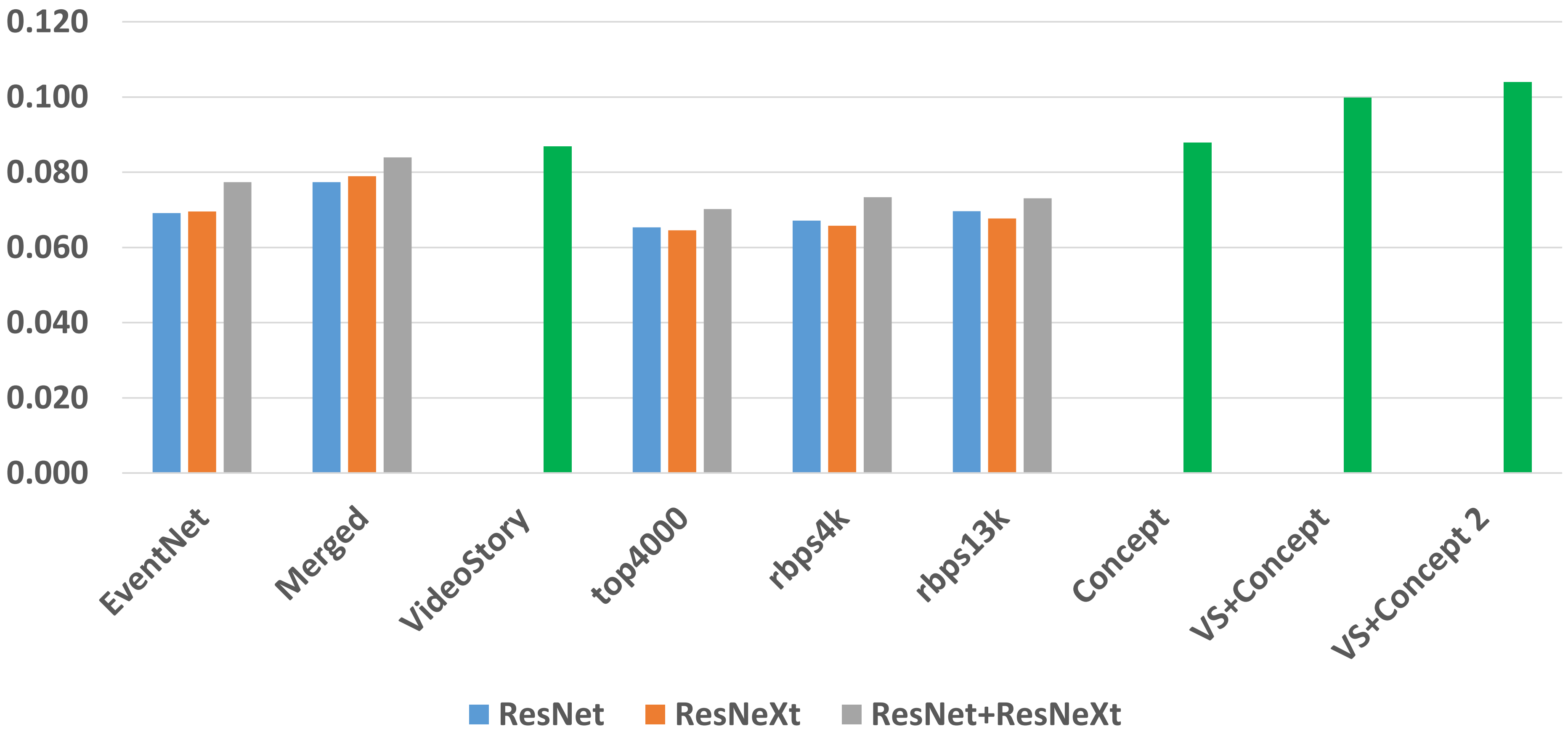
Results 5 Modalities x 2 Features

- VideoStory : ResNeXt is better than ResNet
- Concepts : ResNet is better than ResNeXt (overfit?)
- VideoStory is better than Concepts

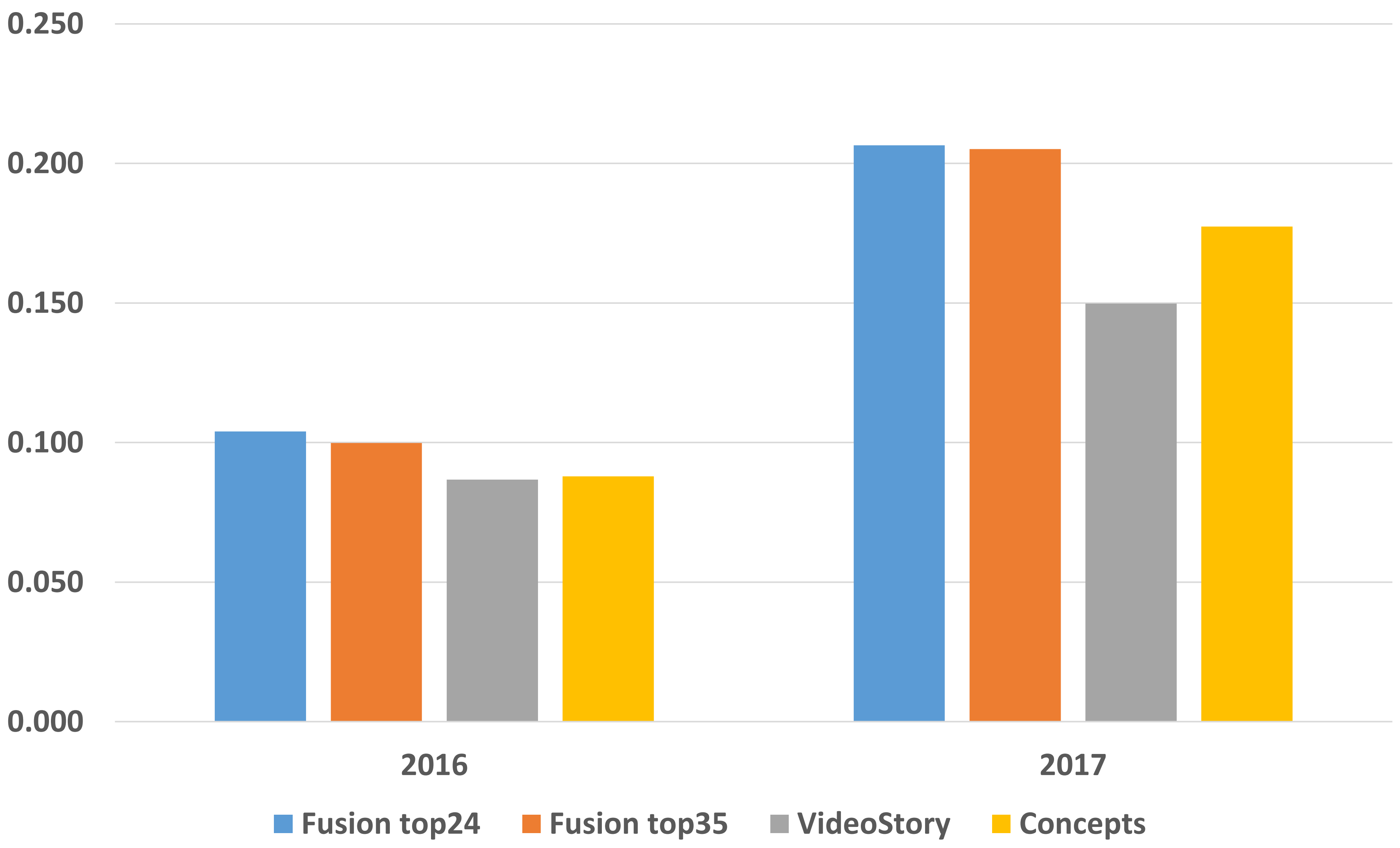


Final Fusion

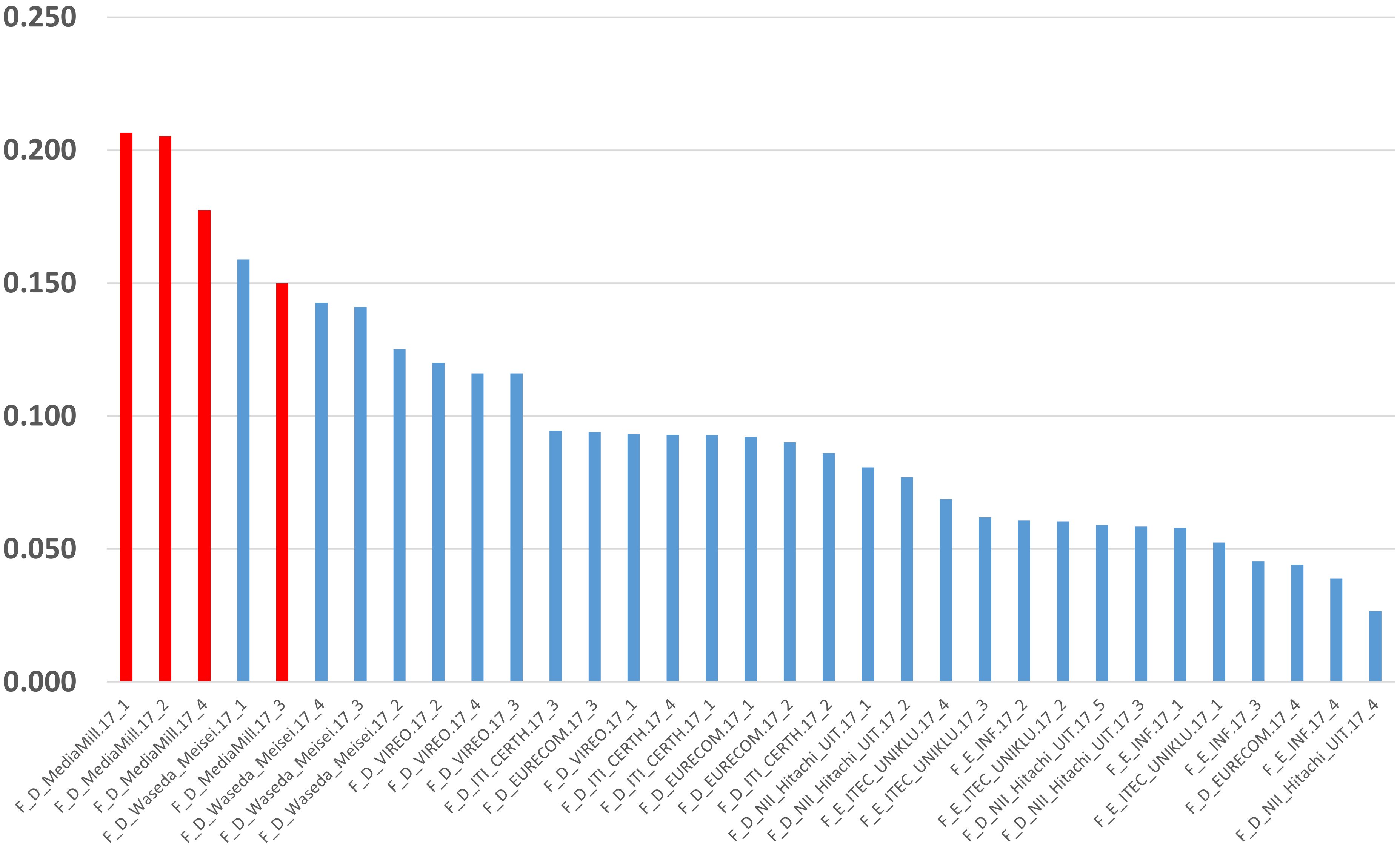
- Concept fusion is slightly better than VideoStory
- Often complementary, also big difference for many topics
- Top 2/4 for concepts is slightly better than top 3/5



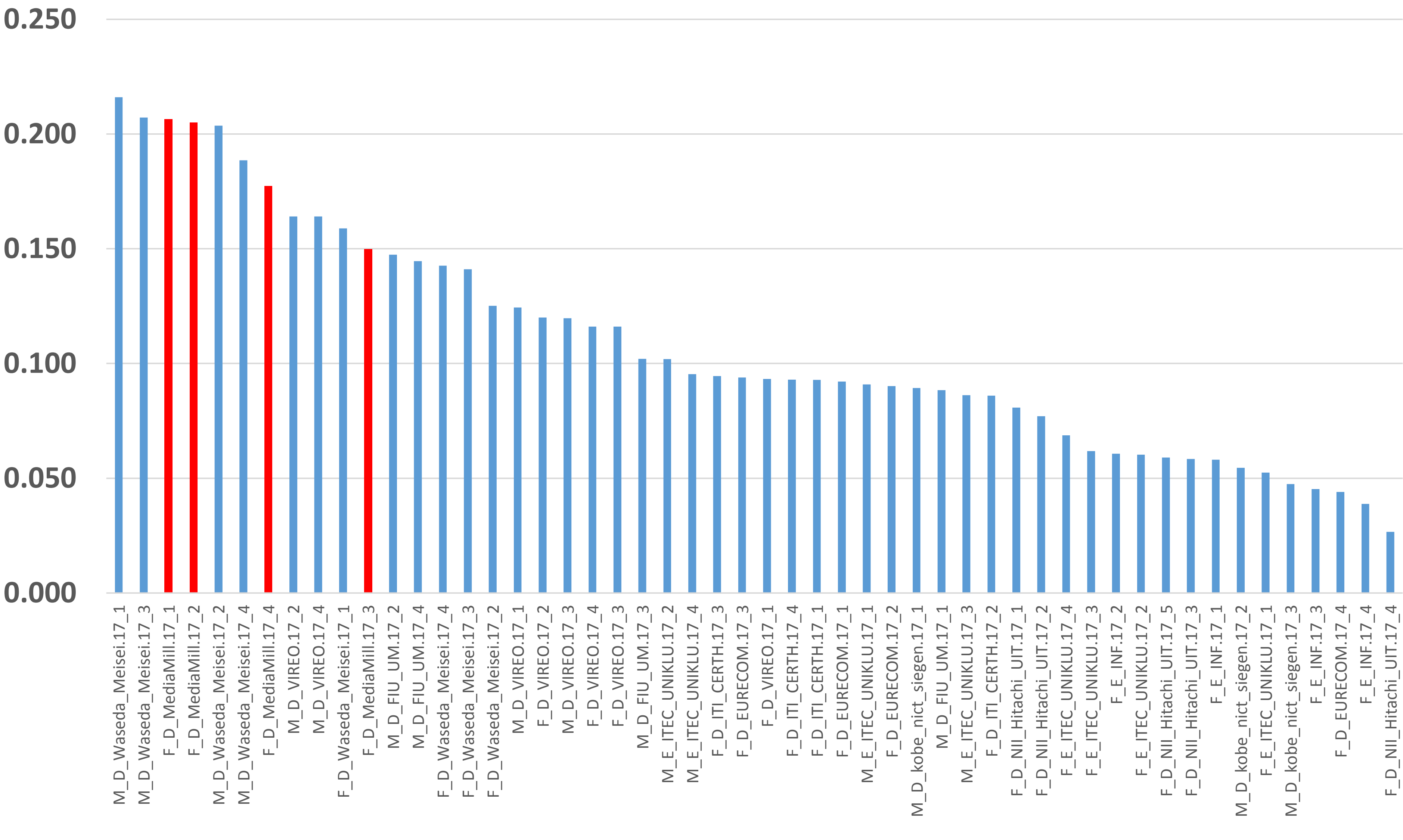
Our AVS Submission



All Fully Automatic AVS Submissions



All Automatic and Interactive AVS Submissions



Conclusions

- Query parsing is important
- VideoStory and Concepts are good but will not “solve” AVS

Thank You