Video to text

Cees Snoek

University of Amsterdam
The Netherlands

NWO

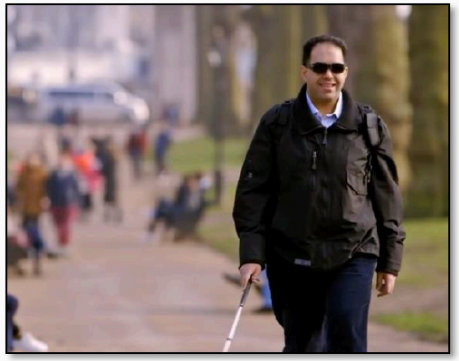
QUVA
Deep Vision Lab

Slides credit

Large majority of slides copied from overview talk at TRECVID 2016 by **Alan Smeaton, Marc Ritter** and **George Awad**.

Why video to text?

robotic vision



assist for blinded

incident report for surveillance

multimedia search

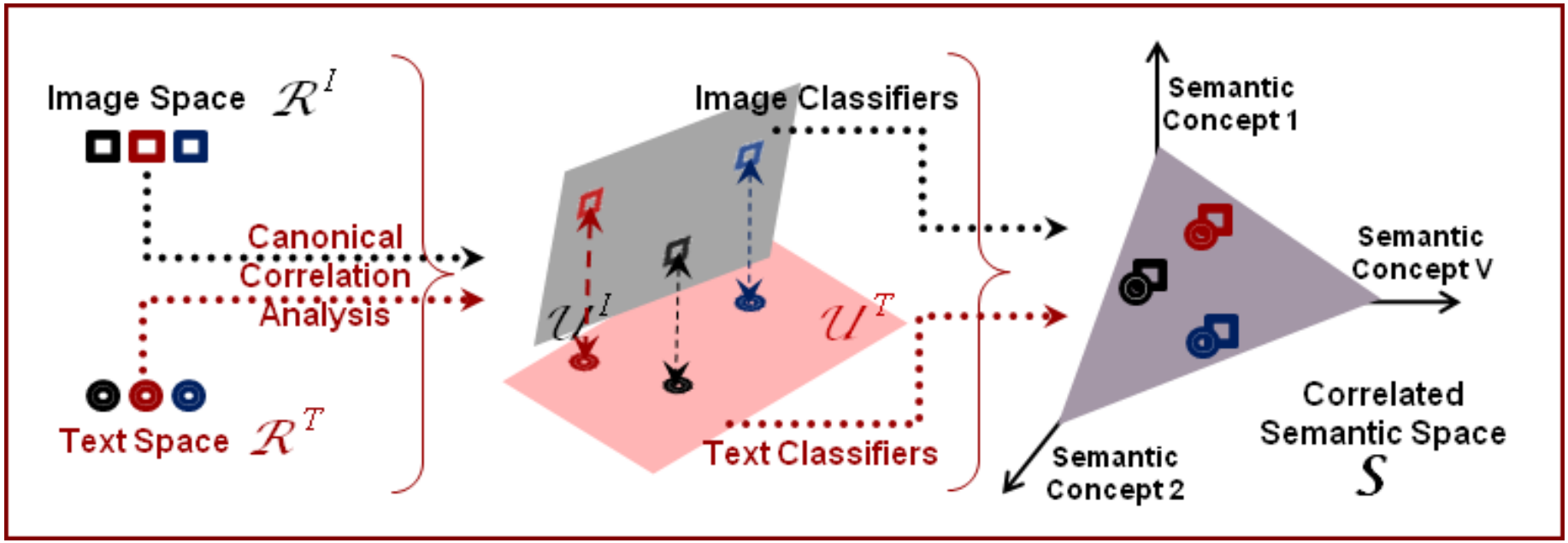
movie description for blinded

seeing chat bot

Slide credit: Tao Mei

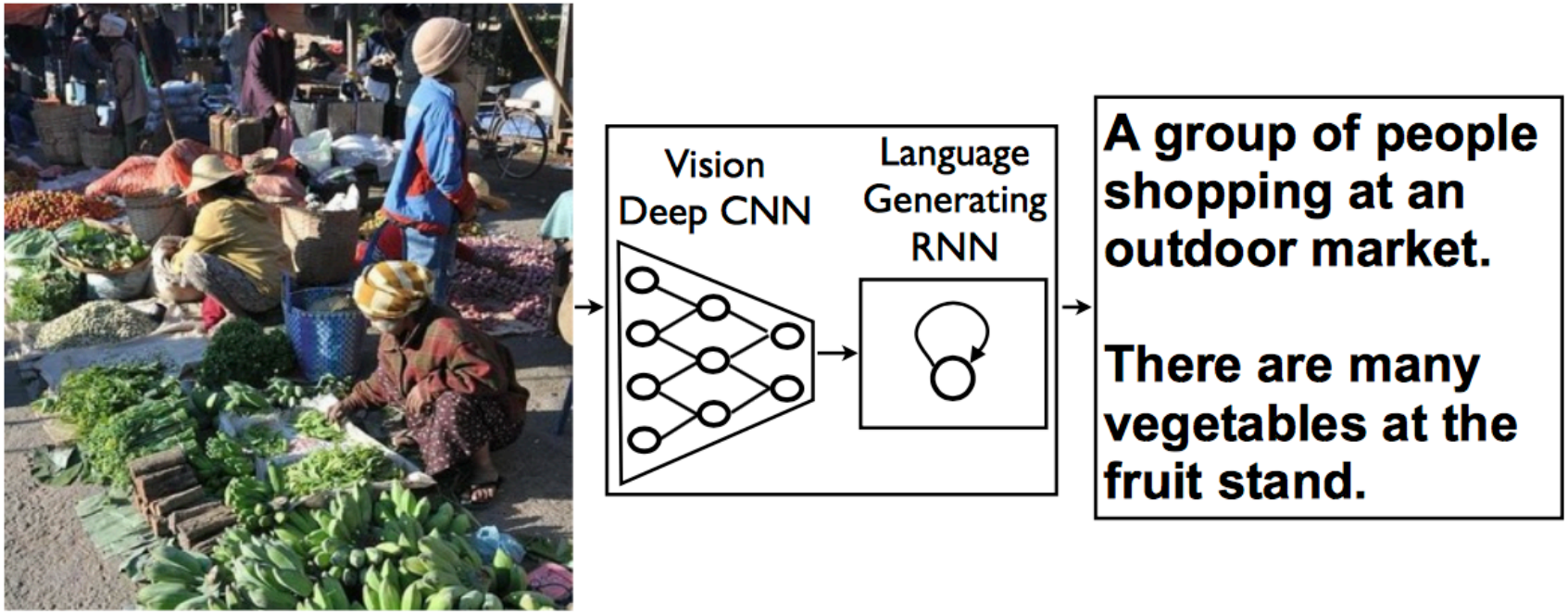
Rasiwasia et al., MM 2010

A typical x-media retrieval baseline



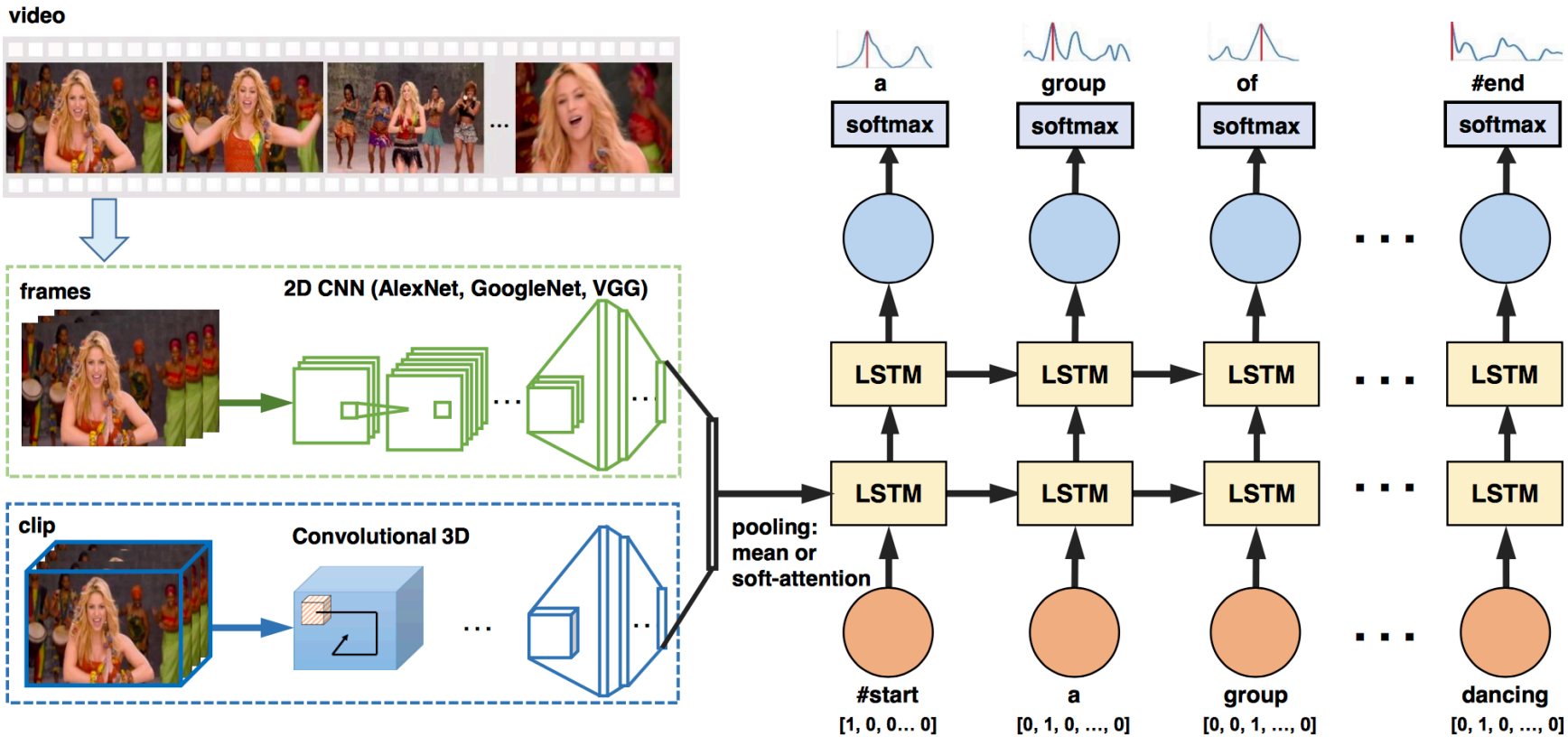
Vinyals et al., CVPR 2015
Karpathy & Fei-Fei, CVPR 2015

A typical image captioning baseline



Xu et al., CVPR 2016

A typical video captioning baseline



This lecture

Overview of this new TRECVID challenge

I. Goals, data and tasks

II. Task I: Matching text to video

III. Task II: Generating text from video

GOALS, TASKS AND DATA

Goals

Measure how well an automatic system can describe a video in natural language.

Measure how well an automatic system can match high-level textual descriptions to low-level video features.

Transfer successful image captioning technology to the video domain.

Tasks

Given a set of

2000 Twitter vine videos.

2 sets (A and B) of text descriptions for each video.

Task I: ***Matching & Ranking***

Return for each video a ranked list of the most likely text description from each set of A and of B.

Task II: ***Description Generation***

Automatically generate a text description for each video.

Dataset Twitter vine videos

Crawled 30k+ Twitter vine video URLs.

Max video duration == 6 sec.

A subset of 2000 URLs randomly selected.

#761
White guy playing the guitar in a room

#387
An Asian young man sitting is eating something yellow

Alterative video & caption datasets

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.

1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of speculators.

Dataset	Organizer	Context	Source	#Video	#Clip	#Sentence	#Word	Vocabulary	Duration (hr)
YouCook	SUNY Buffalo	Cooking	Labeled	88	-	2,668	42,457	2,711	2.3
TACos	MP Institute	cooking	Labeled	123	7,206	18,227	-	-	-
TACos M-L	MP Institute	cooking	Labeled	185	14,105	52,593	-	-	-
M-VAD	UdeM	movie	DVS	92	48,986	55,905	519,933	18,269	84.6
MPII	MP Institute	movie	DVS+Script	94	68,337	68,375	653,467	24,549	73.6
MSVD	MSR	multi-category	AMT workers	-	1,970	70,028	607,339	13,010	5.3
MSR-VTT (10K)	MSRA	20 categories	AMT workers	5,942	10,000	200,000	1,535,917	28,528	38.7
MSR-VTT (20K)	MSRA	20 categories	AMT workers	14,768	20,000	400,000	4,284,032	49,436	87.8

Slide credit: Tao Mei

Annotation effort

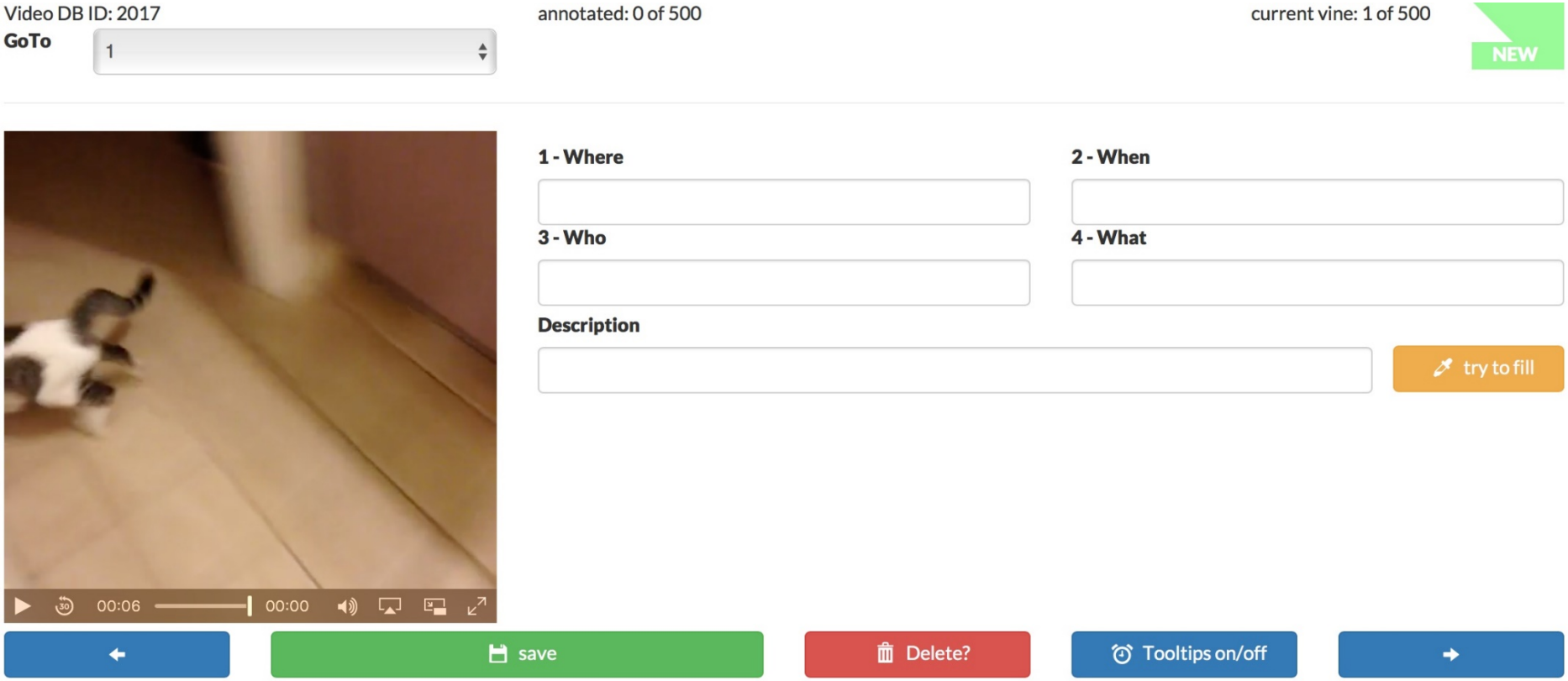
Marc Ritter's TU Chemnitz group supported manual annotation

- Each video annotated by 2 persons (A and B).
- In total 4000 textual descriptions were produced.

NIST asked annotators to combine 4 facets if applicable:

- **Who** is the video describing (objects, persons, animals, …etc) ?
- **What** are the objects and beings doing (actions, states, events, …)?
- **Where** (locale, site, place, geographic, ...etc) ?
- **When** (time of day, season, ...etc) ?

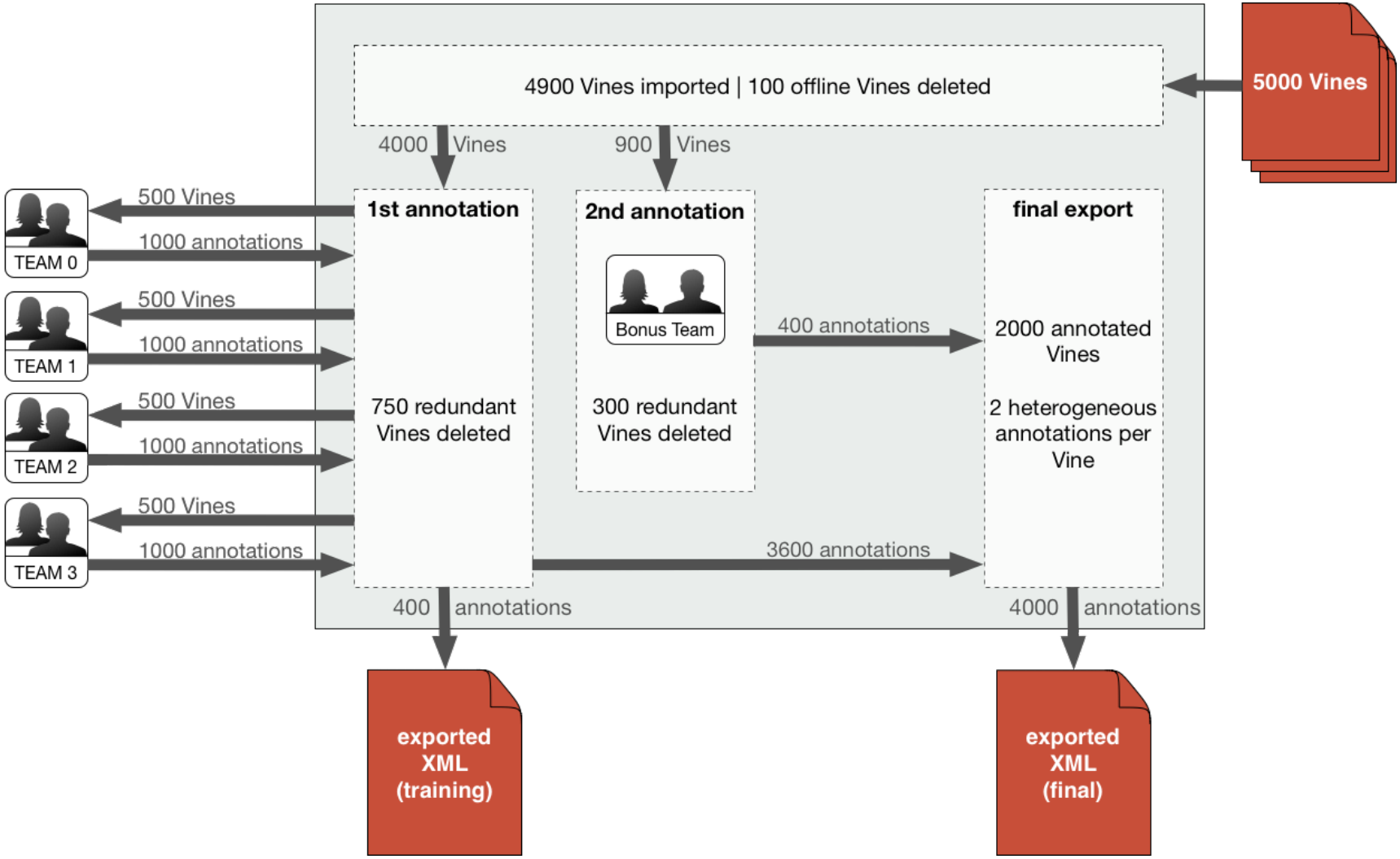
Annotation UI Overview



Annotation challenges

- Bad video quality
- A lot of simple scenes/events with repeating plain descriptions
- A lot of complex scenes containing too many events to be described
- Clips sometimes appear too short for a convenient description
- Audio track relevant for description but has not been used to avoid semantic distractions
- Non-English Text overlays/subtitles hard to understand
- Cultural differences in reception of events/scene content
- Finding a neutral scene description appears as a challenging task
- Well-known people in videos may have influenced (inappropriately) the description of scenes
- Specifying time of day (frequently) impossible for indoor-shots
- Description quality suffers from long annotation hours
- Some offline vines were detected
- A lot of vines with redundant or even identical content

Annotation process



Annotation statistics

UID	# annotations	avg (sec)	max (sec)	min (sec)	# time (hh:mm:ss)
0	700	62.16	239.00	40.00	12:06:12
1	500	84.00	455.00	13.00	11:40:04
2	500	56.84	499.00	09.00	07:53:38
3	500	81.12	491.00	12.00	11:16:00
4	500	234.62	499.00	33.00	32:35:09
5	500	165.38	493.00	30.00	22:58:12
6	500	57.06	333.00	10.00	07:55:32
7	500	64.11	495.00	12.00	08:54:15
8	200	82.14	552.00	68.00	04:33:47
total	**4400**	**98.60**	**552.00**	**09.00**	**119:52:49**

Caption examples

Partition A	Partition B
a dog jumping onto a couch	a dog runs against a couch indoors at daytime
in the daytime, a driver let the steering wheel of car and slip on the slide above his car in the street	on a car on a street the driver climb out of his moving car and use the slide on cargo area of the car
an asian woman turns her head	an asian young woman is yelling at another one that poses to the camera
a woman sings outdoors	a woman walks through a floor at daytime
a person floating in a wind tunnel	a person dances in the air in a wind tunnel

Task 1

MATCHING TEXT TO VIDEO

Task 1: Matching & Ranking

Person reading newspaper outdoors at daytime

Person playing golf outdoors in the field

Three men running in the street at daytime

Two men looking at laptop in an office

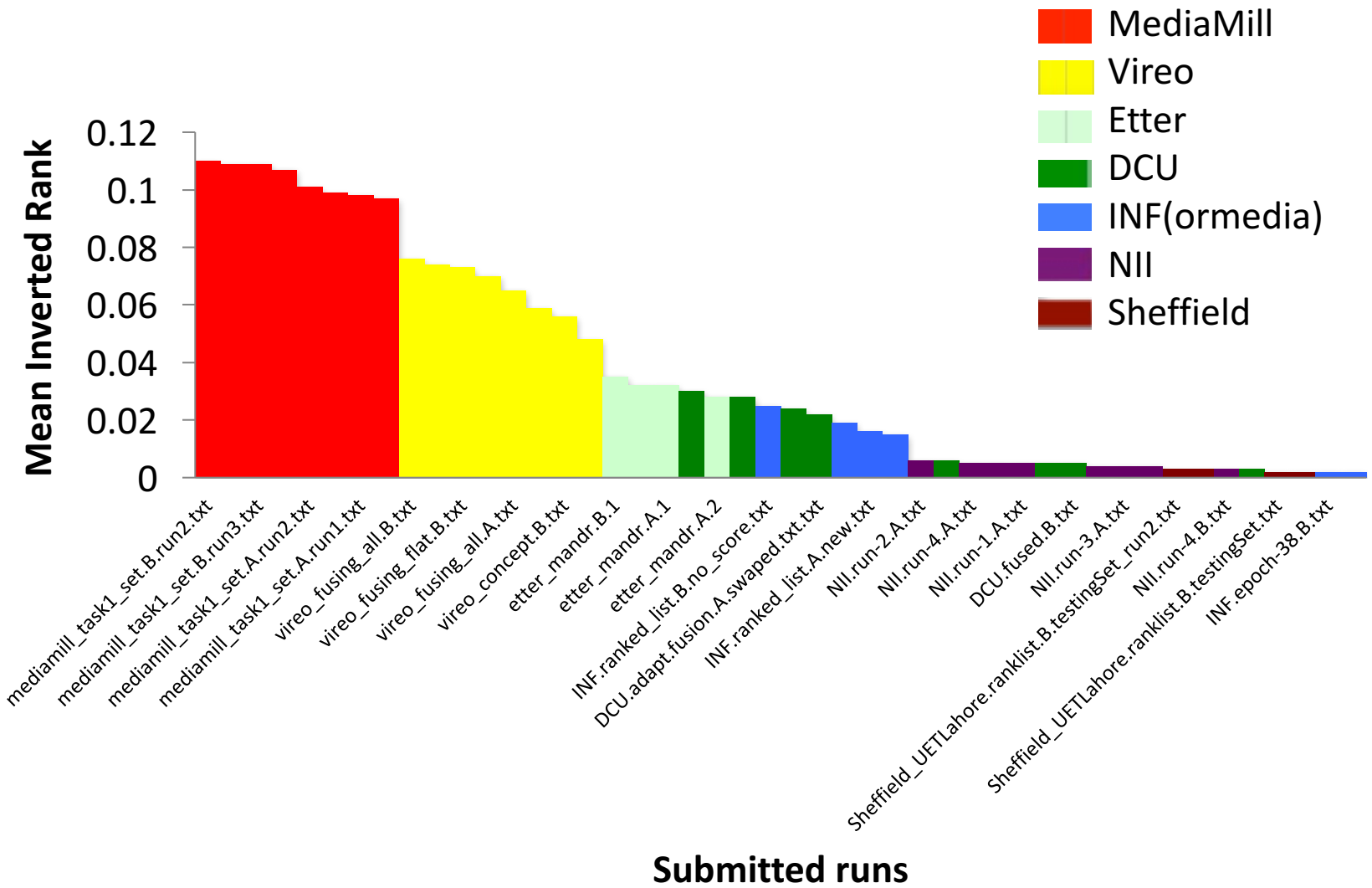
x 2000

x 2000 type A … and … X 2000 type B

Mean inverted rank

The mean inverted rank is the average of the inverted ranks of results for a sample of queries.

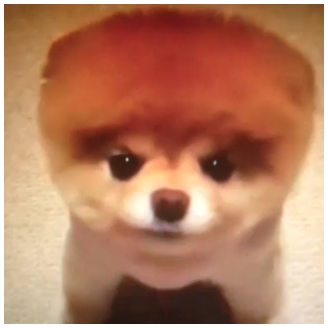
Matching & Ranking results by run



Samples of top 3 results (set A)

#1271
a woman and a man are kissing each other

#1387
a dog imitating a baby by crawling on the floor in a living room



#324
a dog is licking its nose

Samples of bottom 3 results (set B)

#144
A man touches his chin in a tv show

#1060
A man piggybacking another man outdoors

#414
a woman is following a man walking on the street at daytime trying to talk with him

Participant: DCU

Preprocess 10 frames/video to detect 1,000 objects (ImageNet), 94 crowd behaviour concepts (WWW dataset), locations (Places2)

4 runs, baseline BM25, Word2vec, and fusion

Participant: VIREO (CUHK)

Adopted their zero-example MED system in reverse

Used a concept bank of 2000 concepts trained on MSR-VTT, Flickr30k, MS-COCO and TGIF datasets

Runs testing concept-based approach vs attention-based deep models, finding deep models perform better, motion features dominate performance

Participant: Etter Solutions

Focused on concepts for Who, What, When, Where

Used a subset of ImageNet plus scene categories from the Places database

Applied concepts to 1 fps with sliding window, mapped this to document vector, and calculated similarity score

PARTICIPANT: MEDIAMILL

Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. **Word2VisualVec: Image and Video to Sentence Matching by Visual Feature Prediction**. *ArXive*, 2016.

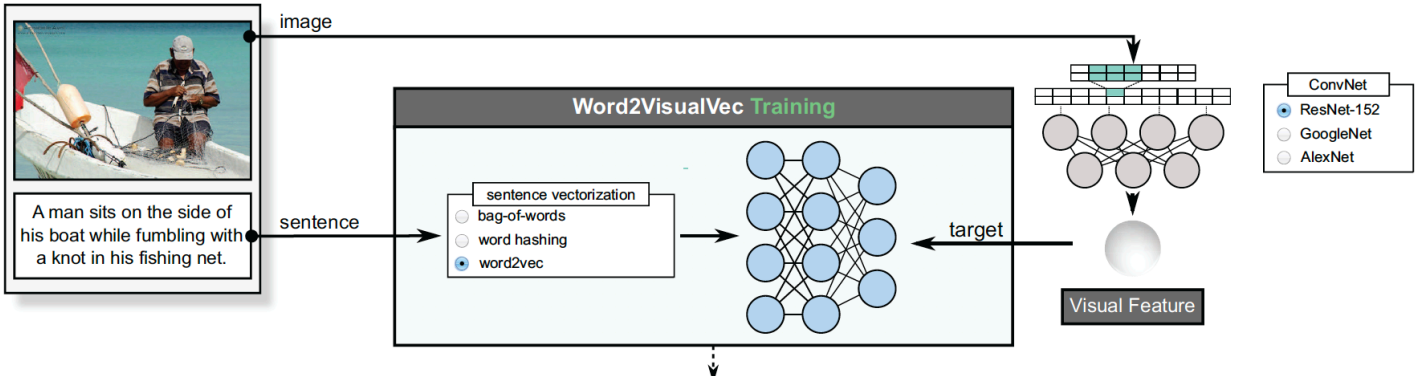
Key idea

Project sentences into a video feature space

Match sentences and videos in this space

Training time

Predict the visual representation of text



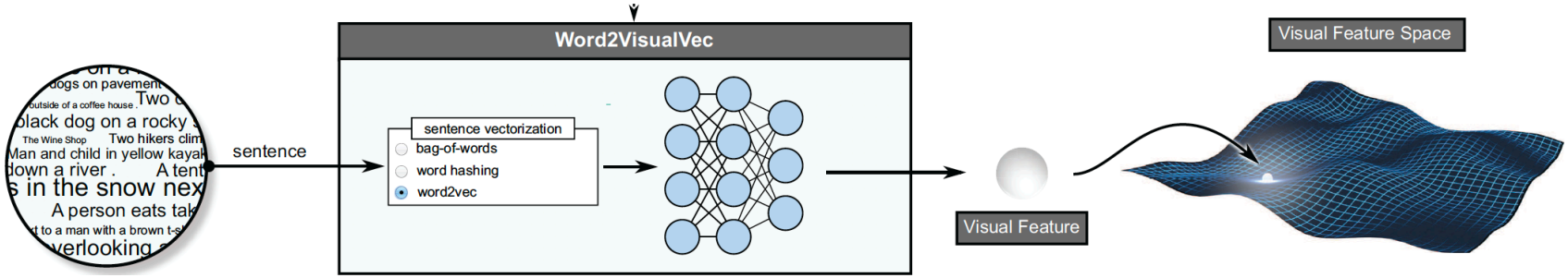
Best settings

Use MSR-VTT for training, minimize mean squared error between text and video
Vine videos for architecture hyperparameter validation

ImageNet-Shuffle with GoogleNet mean pooling over video
Word2vec 500-dim, trained on user tags of 30m Flickr images

Test time

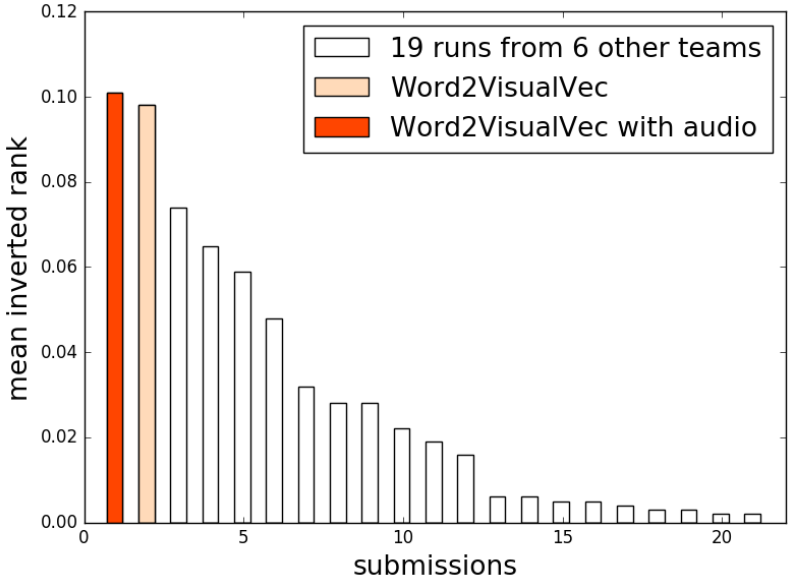
Predict the visual representation of (unseen) text



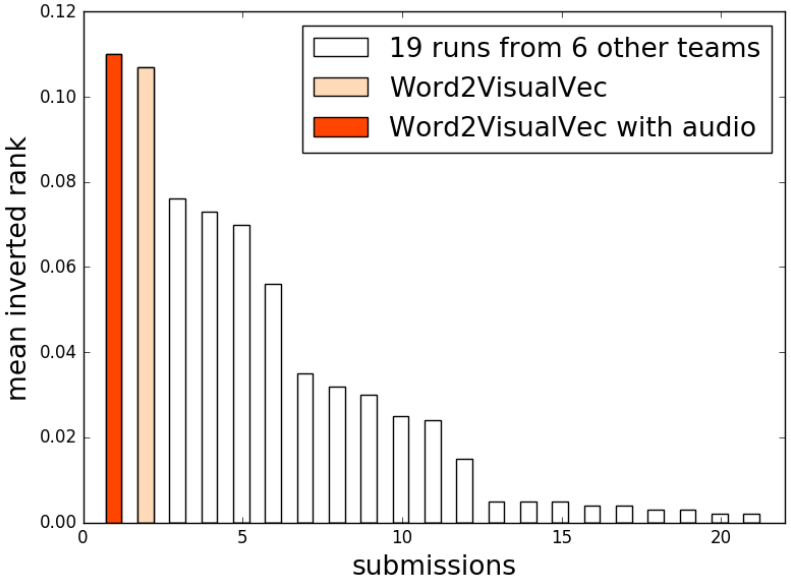
For an unseen video predict its video feature, return closest sentence.

Predicting audio also, small increase

set A



set B



Results

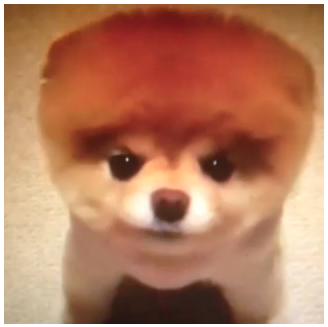

Word2VisualVec: Image and Video to Sentence Matching by Visual Feature Prediction
Jianfeng Dong, Xirong Li, Cees G. M. Snoek
https://arxiv.org/abs/1604.06838

Task 2

GENERATING TEXT FROM VIDEO

Task 2: Description Generation

Given a video



Generate a textual description

Who ? What ? Where ? When ?

"a dog is licking its nose"

BLEU

Approximate human judgement of text quality at corpus level

Measures fraction of common N-grams in source and target.

N-gram matches for a high N rarely occur at sentence-level, so poor performance of BLEU@*N* when comparing only individual sentences, better comparing paragraphs or higher.

Often we see B@1, B@2, B@3, B@4 … we do B@4.

METEOR

Computes unigram precision and recall, extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens

Based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision

Vedantam et al., CVPR 2015

CIDEr

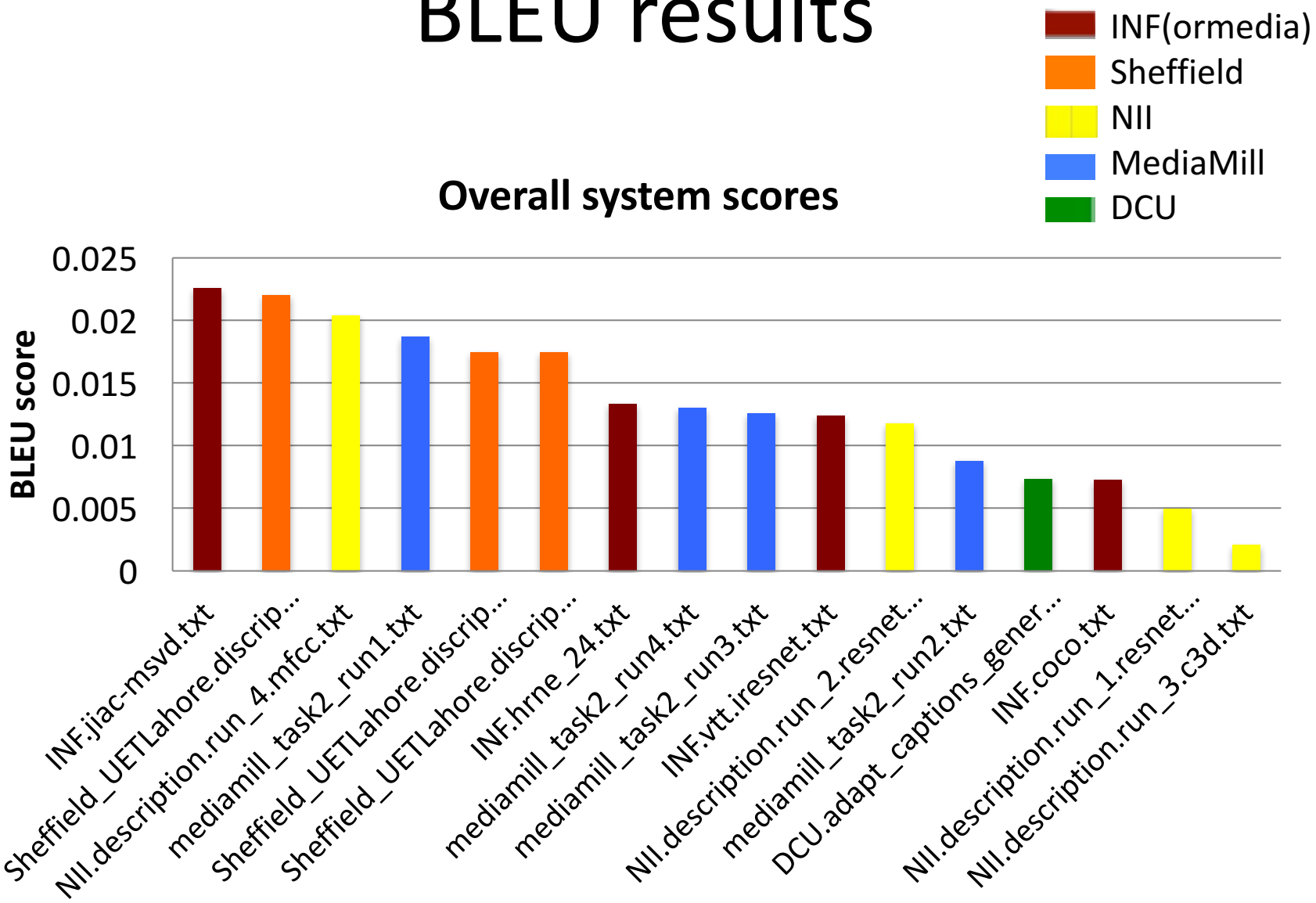
CIDEr (Consensus-Based Image Description Evaluation) is another recent metric

It weights each n-gram with TF-IDF

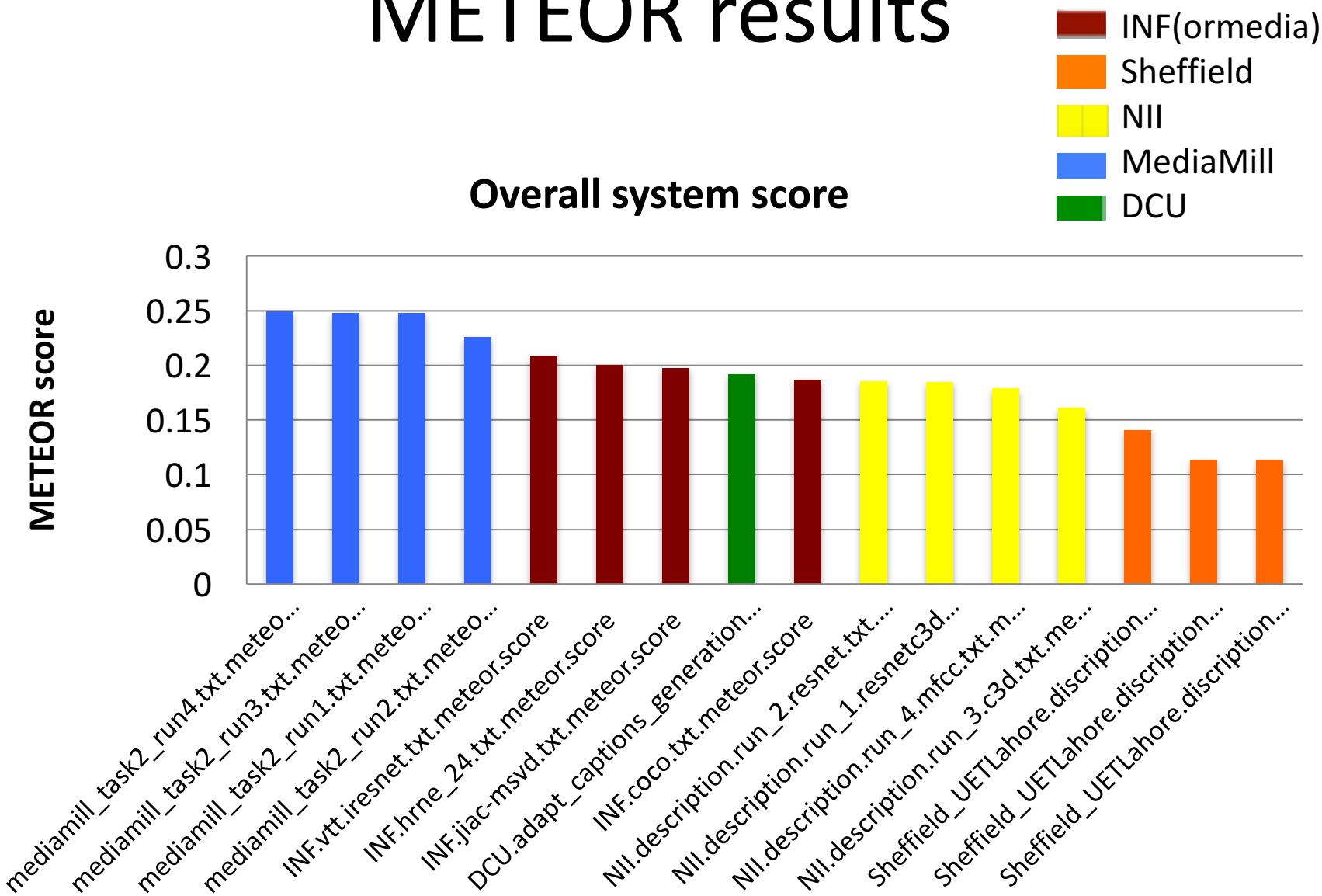
The CIDEr score for n-grams of length n:

- Computed using the average cosine similarity between the candidate and reference sentences

BLEU results



METEOR results



An example from run submissions – 7 unique examples

1. a girl is playing with a baby
2. a little girl is playing with a dog
3. a man is playing with a woman in a room
4. a woman is playing with a baby
5. a man is playing a video game and singing
6. a man is talking to a car
7. A toddler and a dog

Participant: DCU

Train on MS-COCO using NeuralTalk2, a RNN

One caption per keyframe, captions then fused

Participant: Informedia

Focus on generalization ability of caption models

Trained 4 caption models on 3 datasets (MS-COCO, MS-VD, MSR-VTT), achieving sota on those models based on VGGNet concepts and Hierarchical Recurrent Neural Encoder for temporal aspects

Results explore transfer models to TRECVid-VTT

Participant: Sheffield / Lahore

Identified a variety of high level concepts for frames

Detect and recognize faces, age and gender, emotion, objects, (human) actions

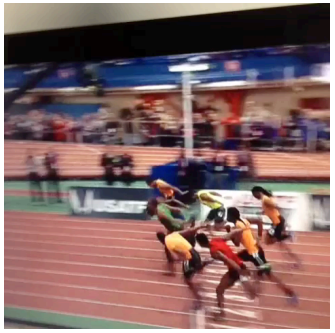
Varied the frequency of frames for each type of recognition

Runs based on combinations of feature types

PARTICIPANT: MEDIAMILL

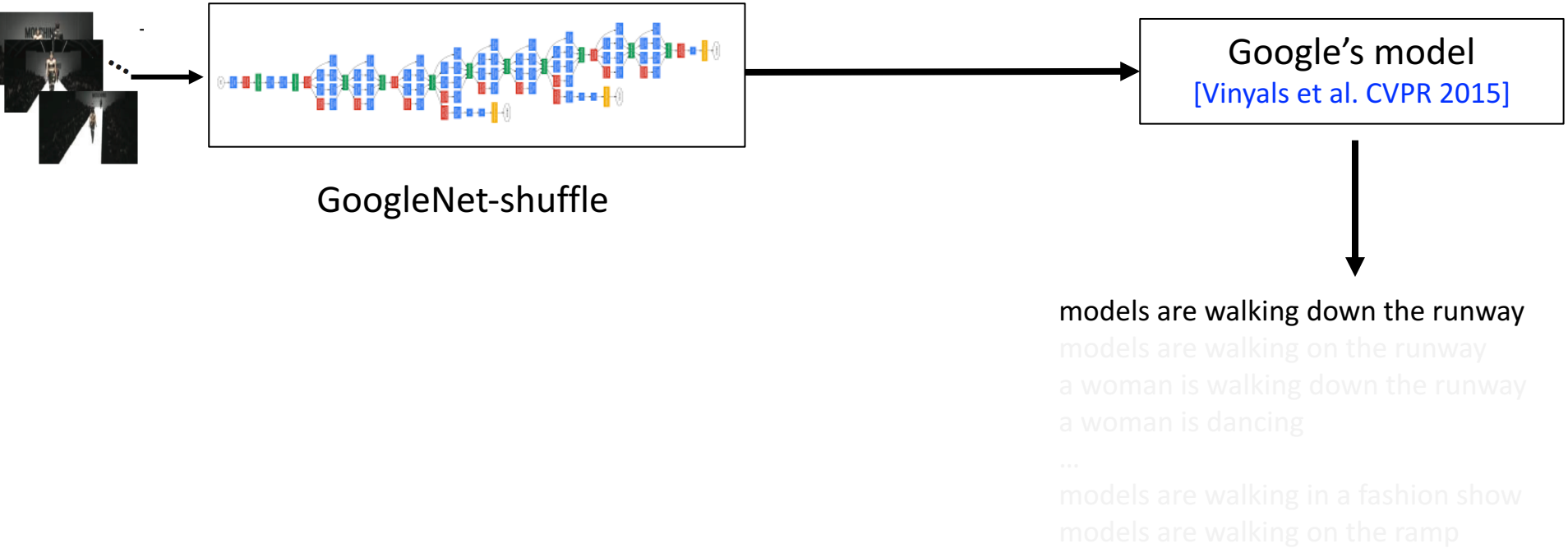
Jianfeng Dong, Xirong Li, Weiyu Lan, Yujia Huo, and Cees G. M. Snoek. **Early Embedding and Late Reranking for Video Captioning**. In *Proceedings of the ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 2016.

Idea: Re-use Video Tags for Captioning

	Predicted tags	Generated caption
	track race field woman	a group of people are running in a **race track**
	soccer player game playing	a **soccer player** is **playing** a goal on a soccer field
	dance people woman dancing	**people** are **dancing** on a stage

Our solution

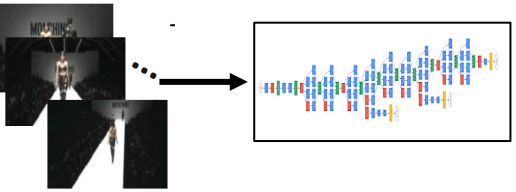
Google's model for sentence generation



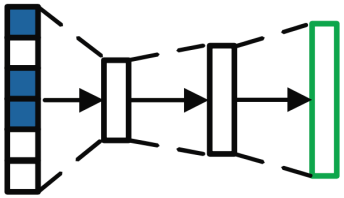
Our solution

Better initialization by tag embedding

Re-encoding by Word2VisualVec



fashion
walking
model

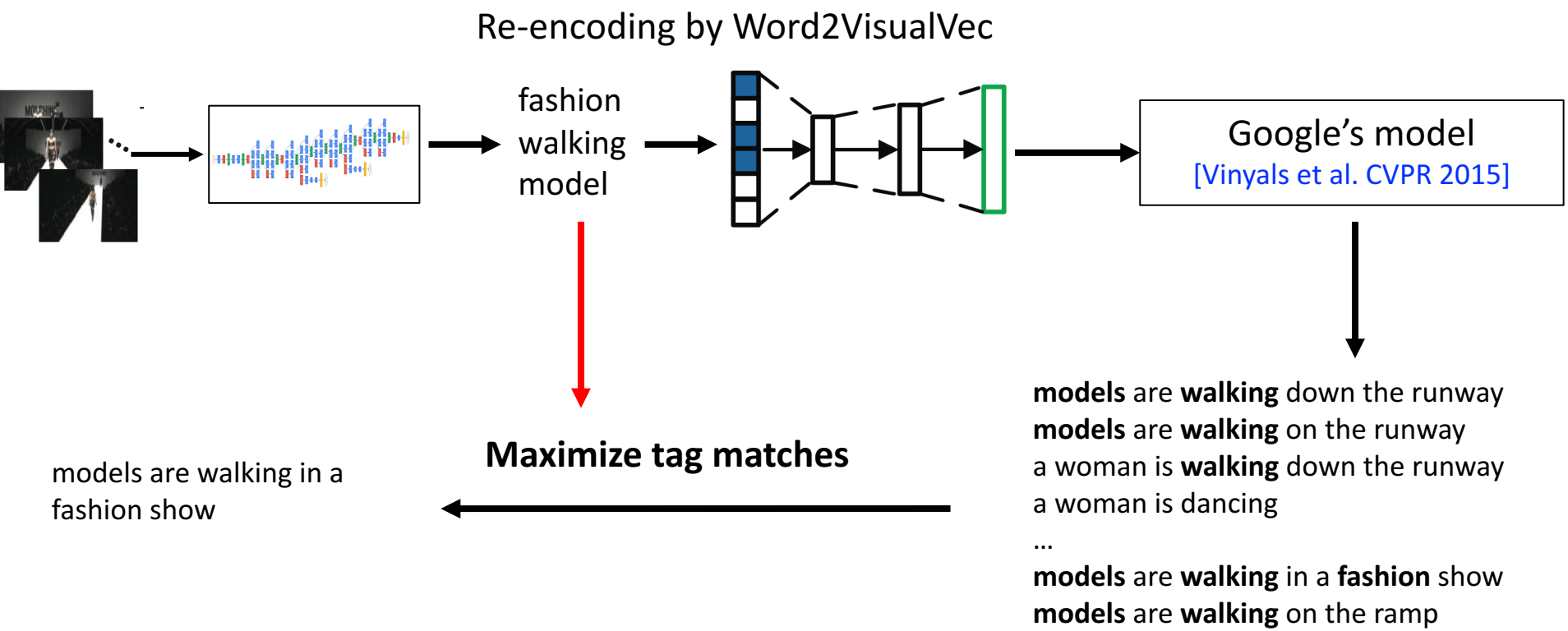


Google's model
[Vinyals et al. CVPR 2015]

models are walking down the runway
models are walking on the runway
a woman is walking down the runway
a woman is dancing
…
models are walking in a fashion show
models are walking on the ramp

Our solution

Rerank sentences by matching with video tags

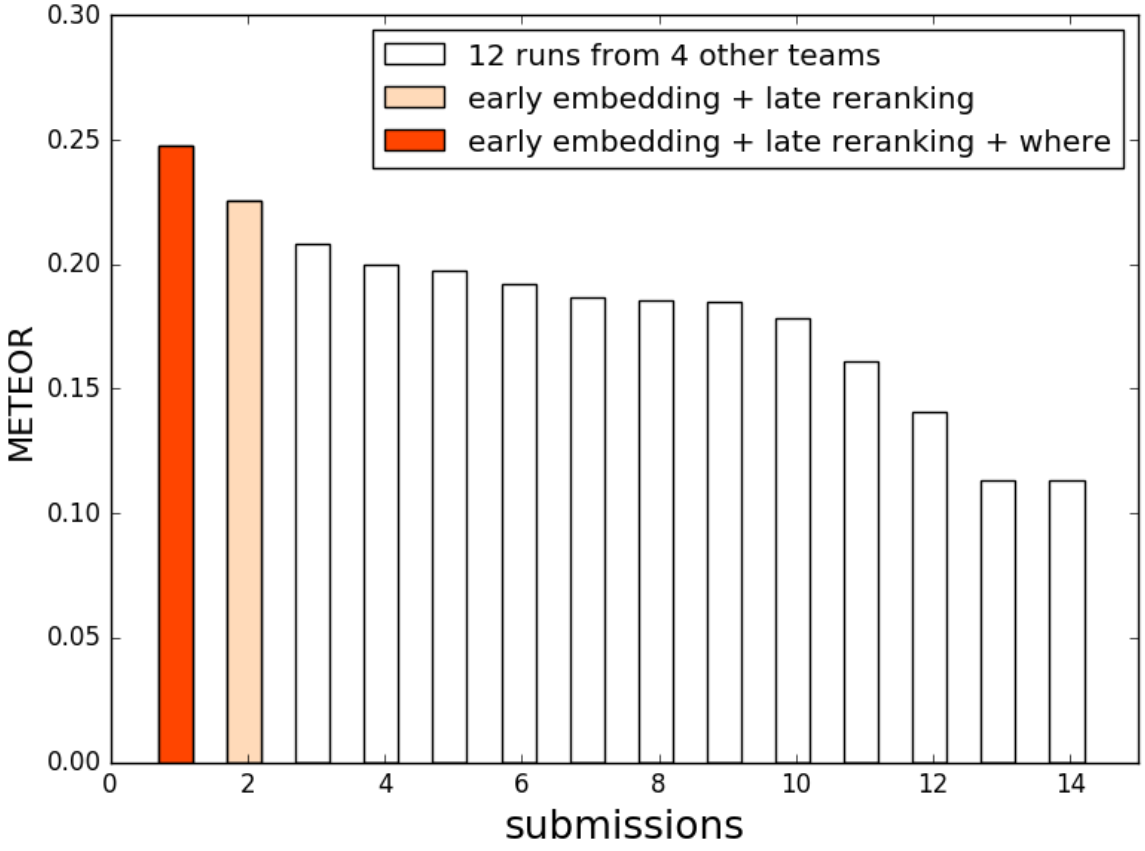


Heuristics to add ‘where’

Two simple rules to append ‘where’ description to the end of the generated sentences:

1. Add “on a $sport_name filed” if $sport appear in the sentence, such as basketball, baseball, and football.
2. Add “on a stage” if “sing” or “dance” appear in the sentence.

Adding “where” improves performance slightly



Conclusions on video to text task

Good participation, good finishing

METEOR scores are higher than BLEU, should have used CIDEr also

Lots of available training sets, some overlap ... MSR-VTT, MS-COCO, Places2, ImageNet, YouTube2Text, MS-VD ..

Only few applied the 4-facet description in their submissions

Task continues in 2017