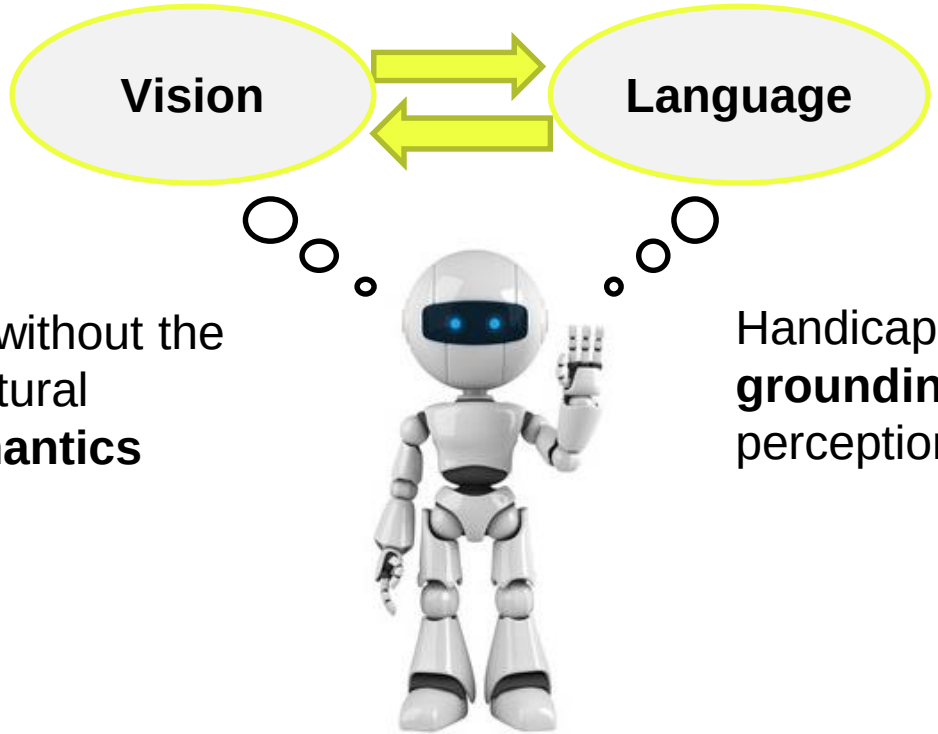
Multimodal Video Description and Caption-Guided Visual Saliency

Vasili Ramanishka

BOSTON UNIVERSITY

Why combine vision and language?



Handicapped without the richness of natural language **semantics**

Handicapped without **grounding** meaning in perception

Why combine vision and language?

IMAGES

CNN

FLOWER
FIELD
SKY
CLOUDS

LANGUAGE

It is a bright summer day, the weather is fine with picturesque clouds over the horizon. Sunflowers are growing in a field under the beautiful blue sky.

Applications:

- Social media analysis
- Security and surveillance
- AI assistants
- Summarization and retrieval
- etc…

How can we connect vision and language?

Tasks:

Captioning Hendricks et al, CVPR16 Ramanishka et al, ACMM16

A crowd of people is looking at giraffes in a zoo.

Referring Expressions Hu et al., CVPR16

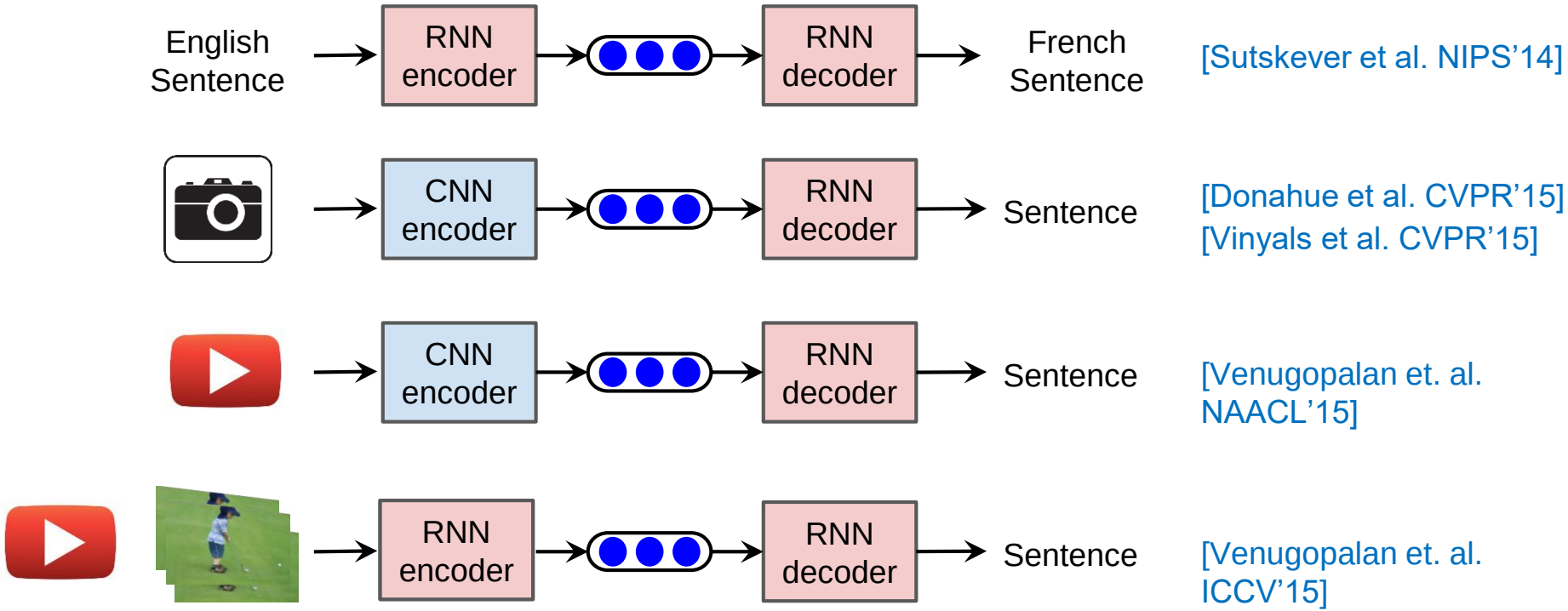
Person taking a photo?

Question Answering Xu and Saenko ECCV16

What time of year is it?

Answer: summer

Encoder-decoder framework



MSR-VTT dataset

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 man and a woman performing a musical.
2. A teenage couple perform in an amateur musical.
3. Dancers are playing a routine.
4. People are dancing in a musical.
5. Some people are acting and singing for performance.

1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.

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.

Xu et al., CVPR 2016

Video description

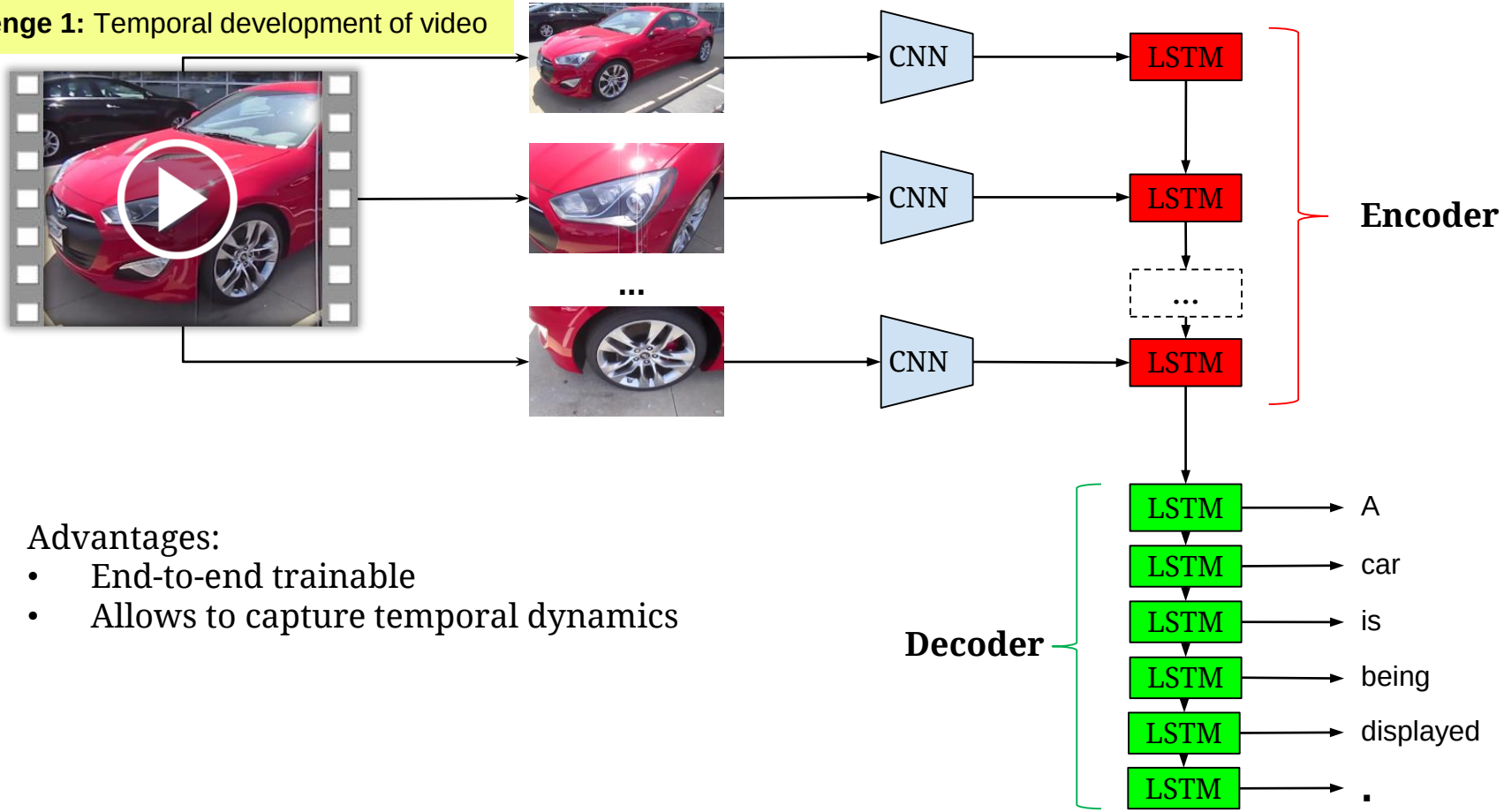
1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

Xu et al., CVPR 2016

Problems to be addressed:

- Temporal development of video
- Capture activities and small motion
- Capture information from audio
- Topic-aware model to capture language nuances

Challenge 1: Temporal development of video

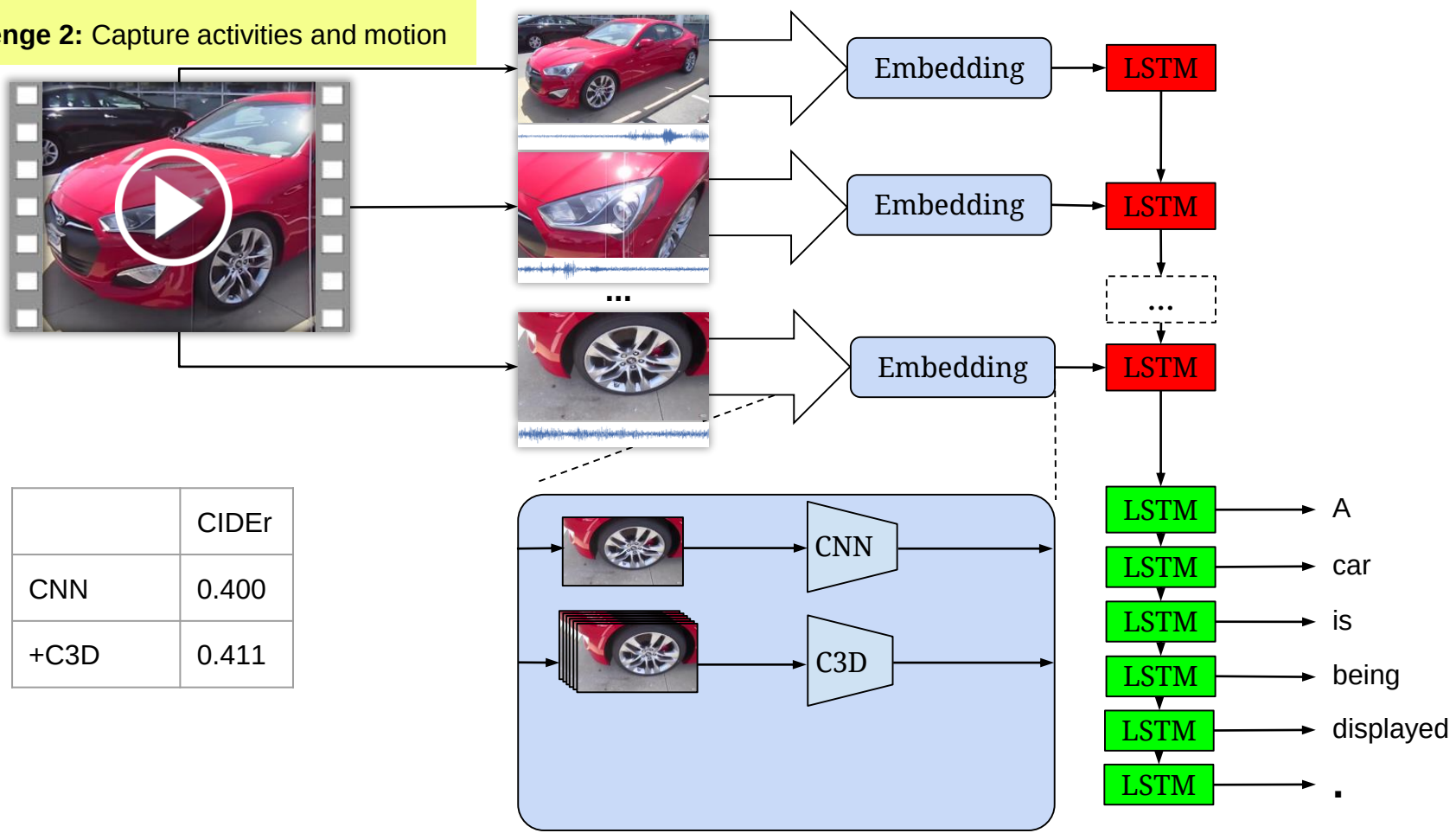


Advantages:

- End-to-end trainable
- Allows to capture temporal dynamics

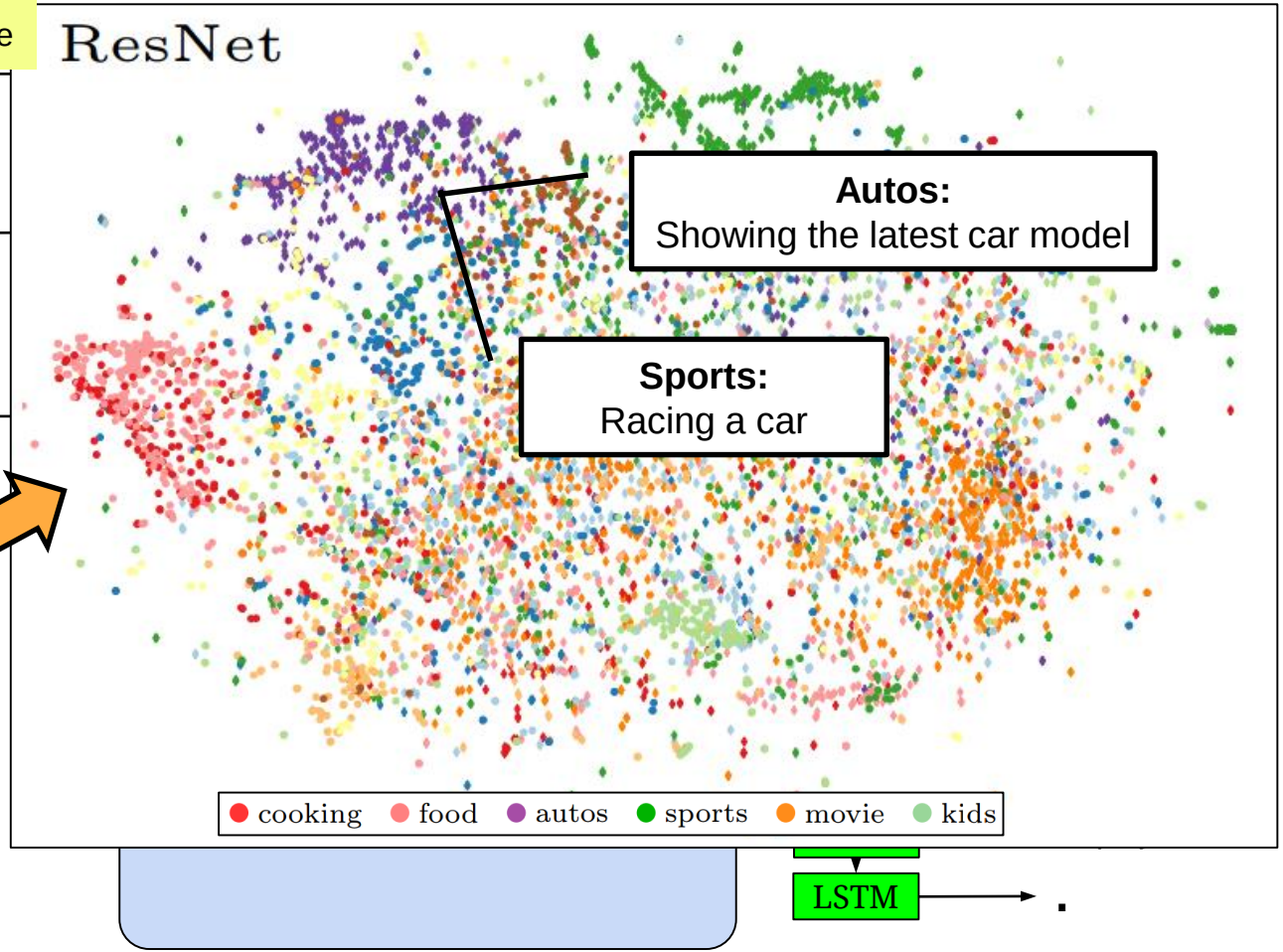
Venugopalan et al., ICCV 2015

Challenge 2: Capture activities and motion



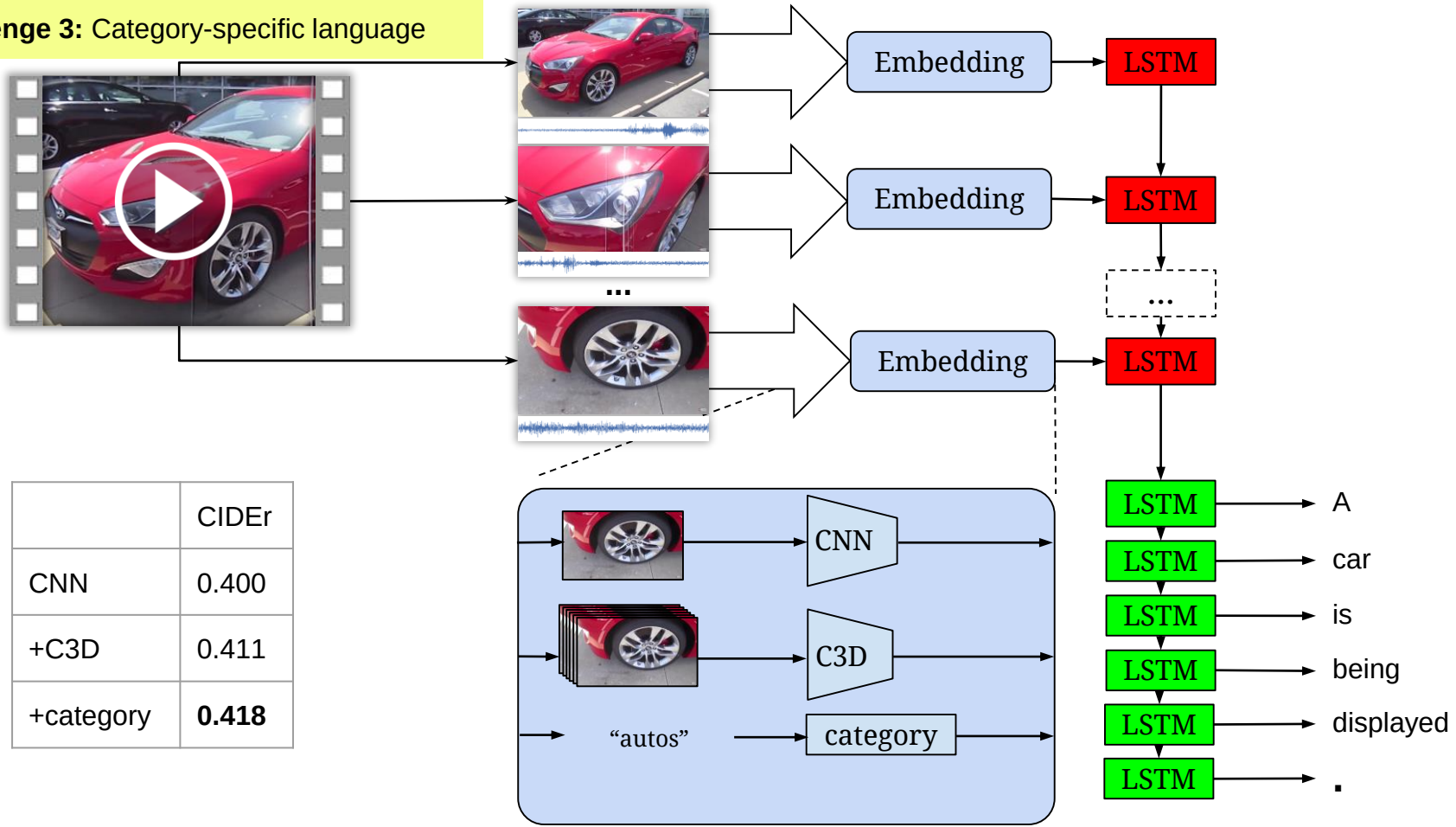
	CIDEr
CNN	0.400
+C3D	0.411

Challenge 3: Category-specific language



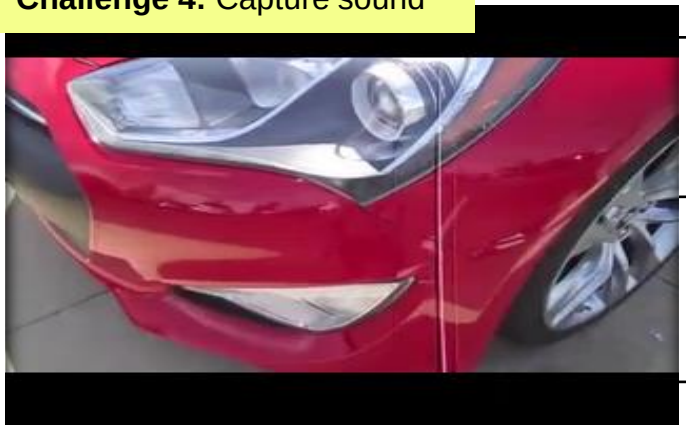
	CIDEr
CNN	0.400
+C3D	0.411

Challenge 3: Category-specific language



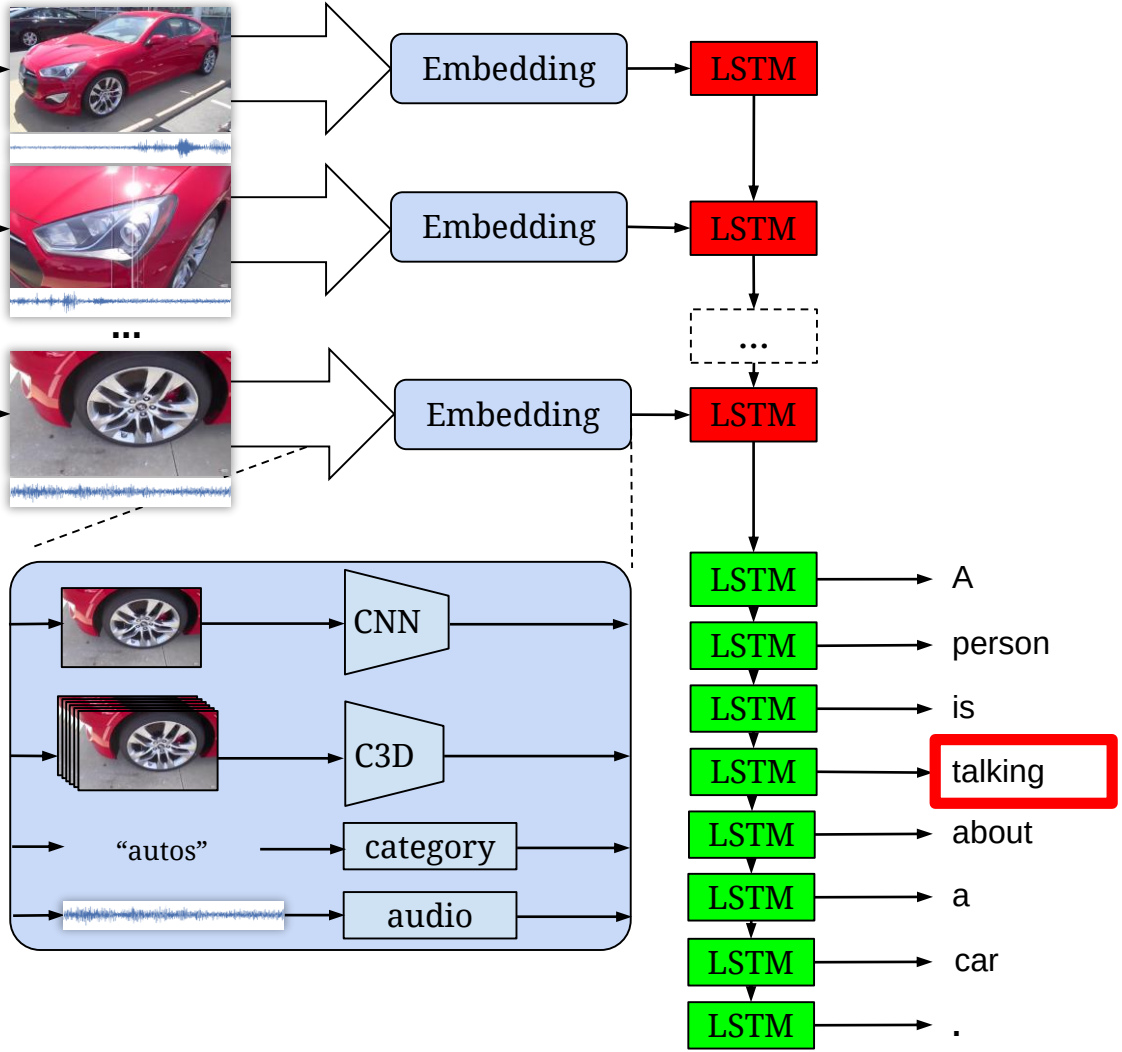
	CIDEr
CNN	0.400
+C3D	0.411
+category	**0.418**

Challenge 4: Capture sound

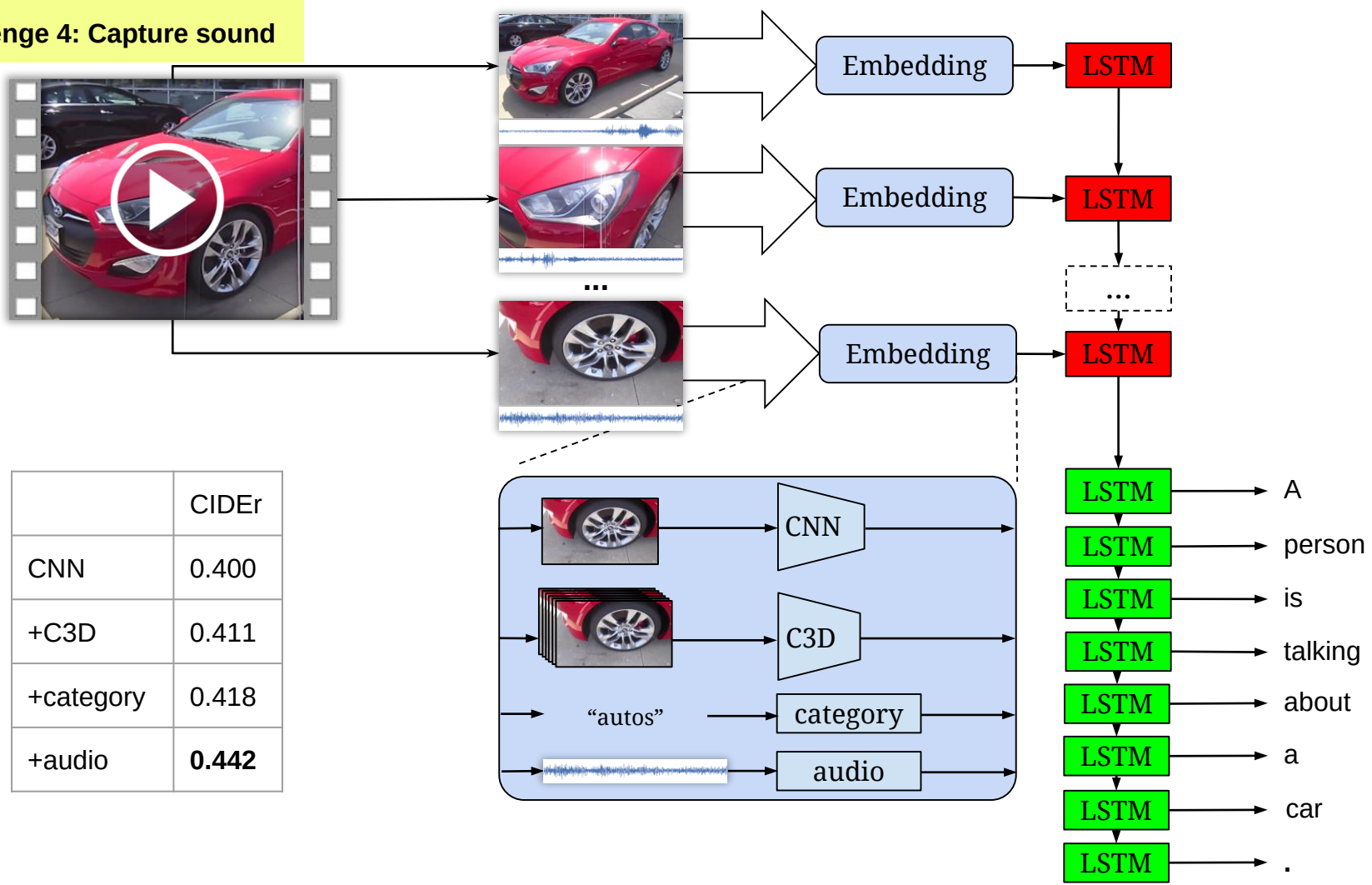


With audio

	CIDEr
CNN	0.400
+C3D	0.411
+category	0.418

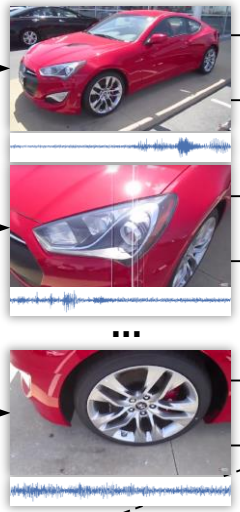


Challenge 4: Capture sound



	CIDEr
CNN	0.400
+C3D	0.411
+category	0.418
+audio	**0.442**

Challenge 5: Factor language model



Embedding

Embedding

Embedding

LSTM

LSTM

...

LSTM

LSTM

LSTM

LSTM

LSTM

LSTM

LSTM

LSTM

LSTM

A

person

is

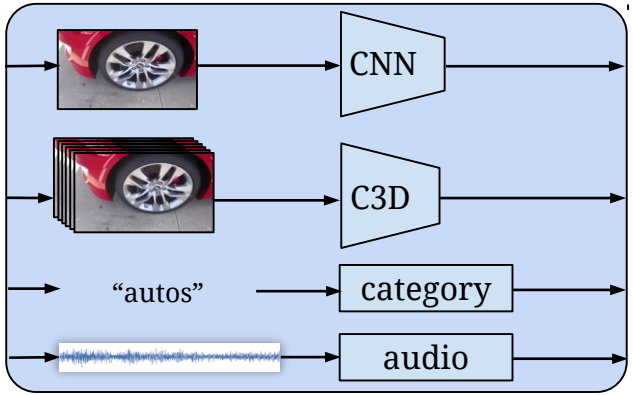
talking

about

a

car

.

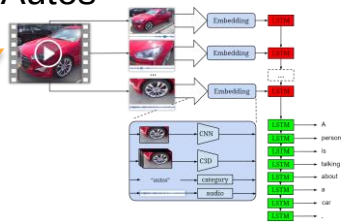


Science: "man", "talking"

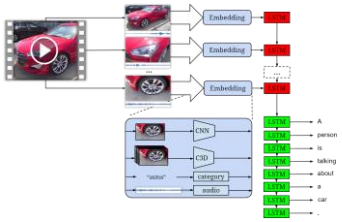
Cooking: "woman", "cooking"

Challenge 5: Factor language model

Autos

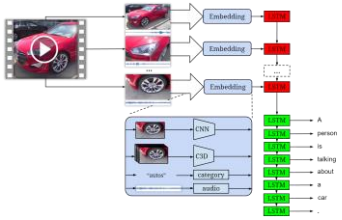


Science



...

Cooking



Reference descriptions:

"A man is talking about a car"

"A narrator speaks over a promotional video from a car manufacturer about the innovations the manufacturer has made to cars"

"A grill that attaches to the back of a car is shown"

	CIDEr
CNN	0.400
+C3D	0.411
+category	0.418
+audio	0.442
experts	**0.465**

Network of experts

	CIDEr
audio	0.184
categories	0.236
C3D	0.389
CNN	0.400
+C3D	0.411
+category	0.418
+audio	0.442
experts	**0.465**

Summary

- Temporal development of video
  -> Encoder – Decoder approach (S2VT)
- Capture activities and motion
  -> C3D representation extracted from 16 frame batches
- Capture sound and audio
  -> MFCC as audio features
- Topic-aware model to capture language differences
  -> Network of experts

ACM MM 2016 Video Description Challenge

Human evaluation

Rank	Team	Organization	**Coherence**	**Relevance**	**Helpful for blind**
1	Aalto	Aalto University	3.263	3.104	3.244
2	v2t_navigator	RUC & CMU	3.261	3.091	3.154
3	**VideoLAB**	UML & Berkeley & UT-Austin	3.237	**3.109**	3.143
...					
21					

ACM MM 2016 Video Description Challenge

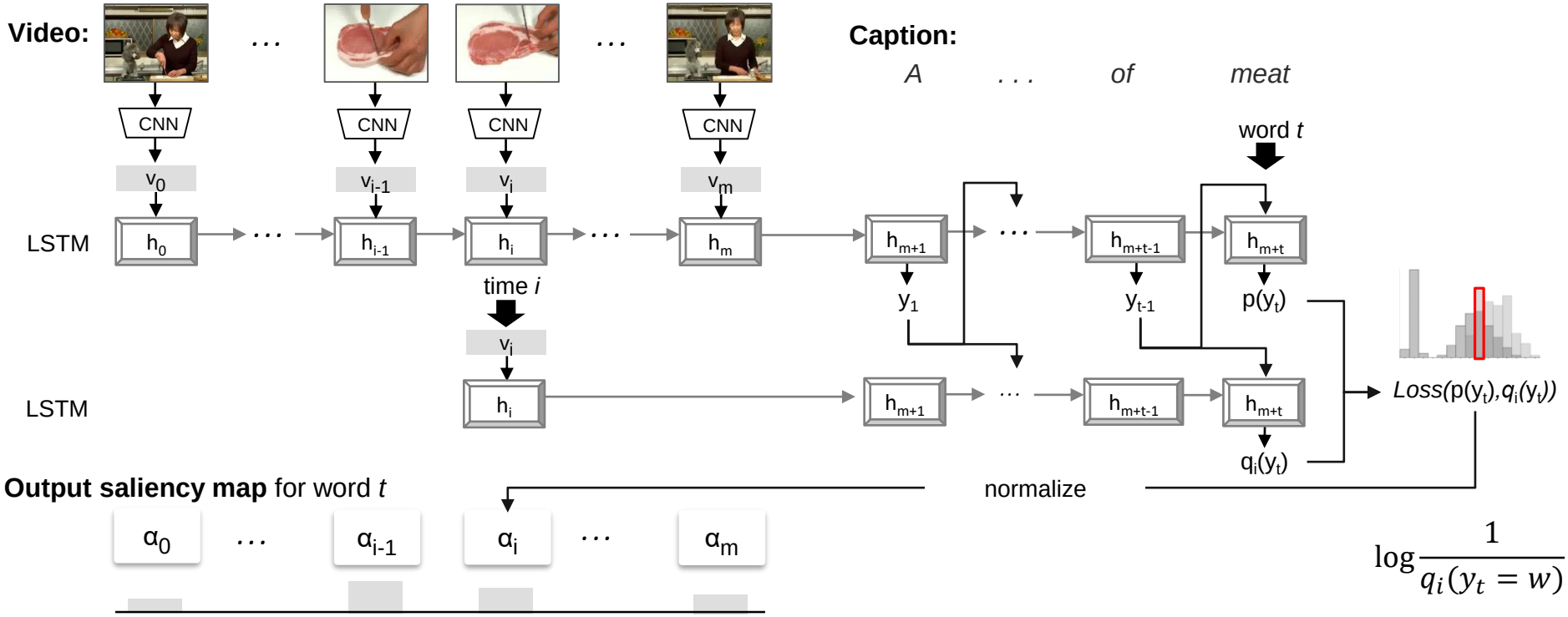
Automatic evaluation

Rank	Team	Organization	BLEU@4	Meteor	CIDEr-D	ROUGE-L
1	v2t_navigator	RUC & CMU	0.408	0.282	0.448	0.609
2	Aalto	Aalto University	0.398	0.269	0.457	0.598
3	**VideoLAB**	UML & Berkeley & UT-Austin	0.391	0.277	0.441	0.606
…						
21						

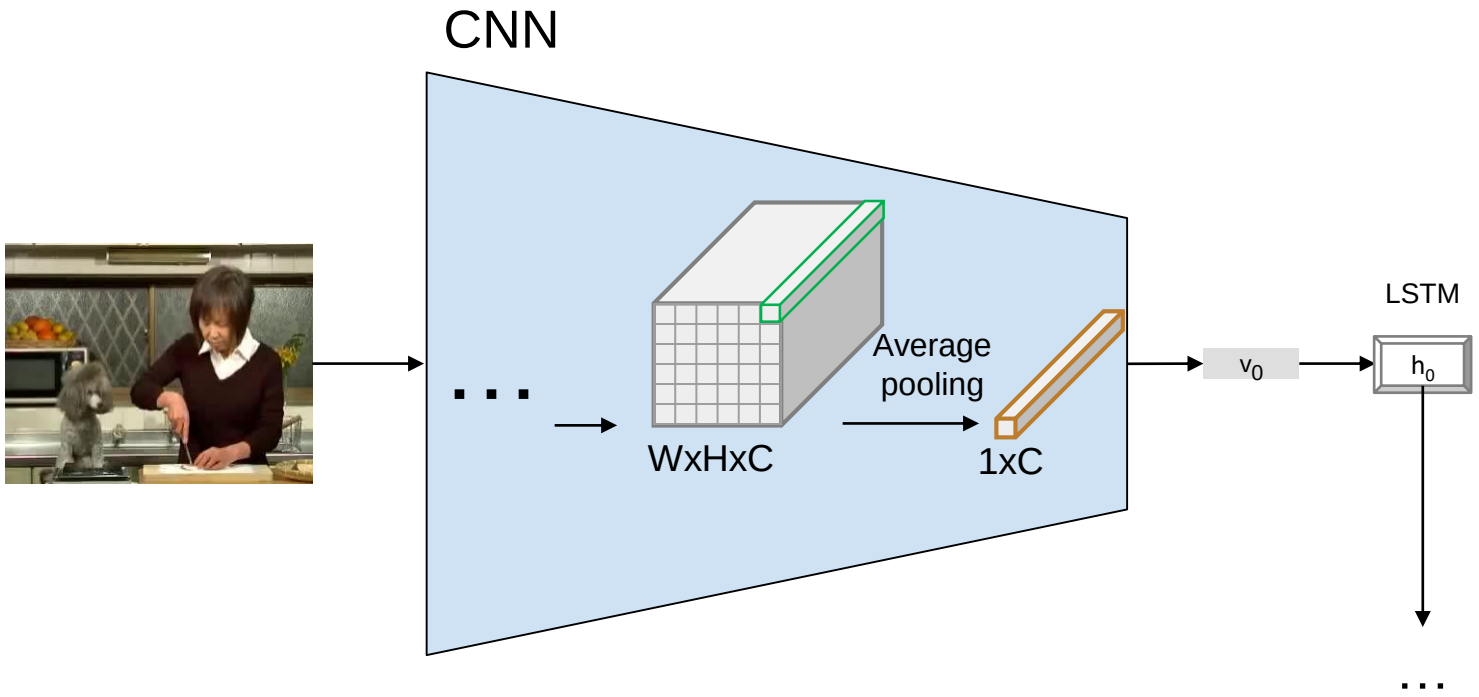
Visual Saliency

Predicted sentence: A woman is cutting a piece of meat

Approach

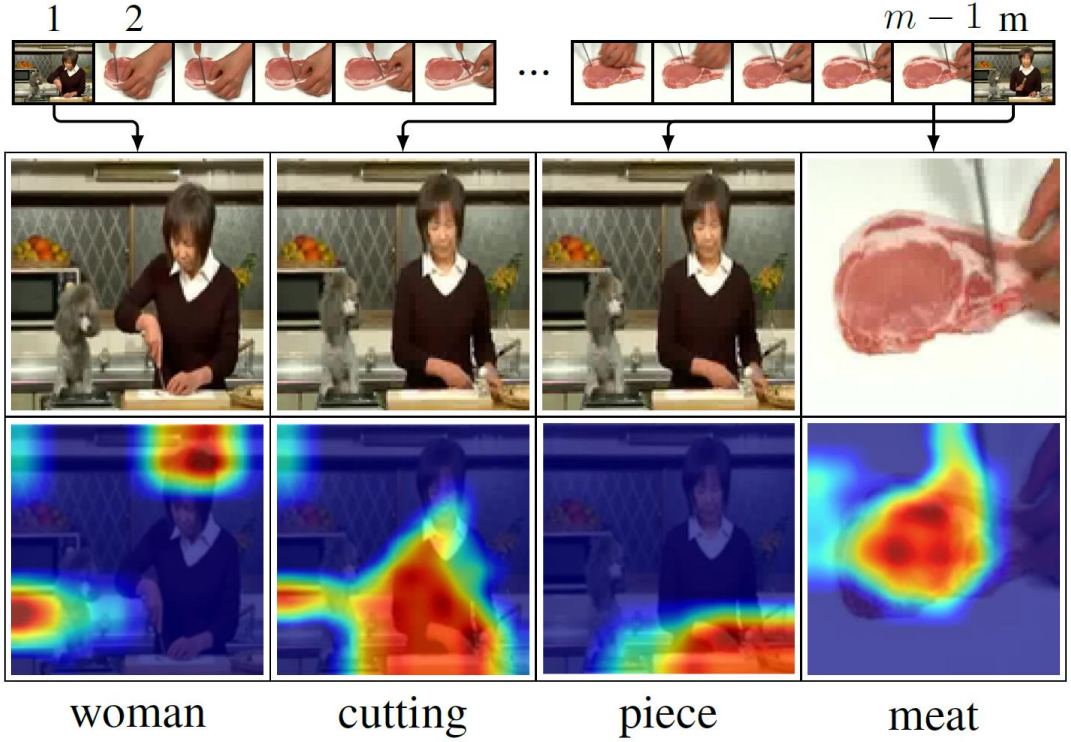

Video:
Caption:
A . . . of meat
CNN
v0
vi-1
vi
vm
LSTM
h0
hi-1
hi
hm
time i
hm+1
hm+t-1
hm+t
word t
y1
yt-1
p(yt)
qi(yt)
Loss(p(yt),qi(yt))
normalize
Output saliency map for word t
α0
αi-1
αi
αm
$\log \frac{1}{q_i(y_t = w)}$

Spatial localization (almost) for free



Spatiotemporal saliency

Predicted sentence: A **woman** is **cutting** a **piece** of **meat**



Spatiotemporal saliency

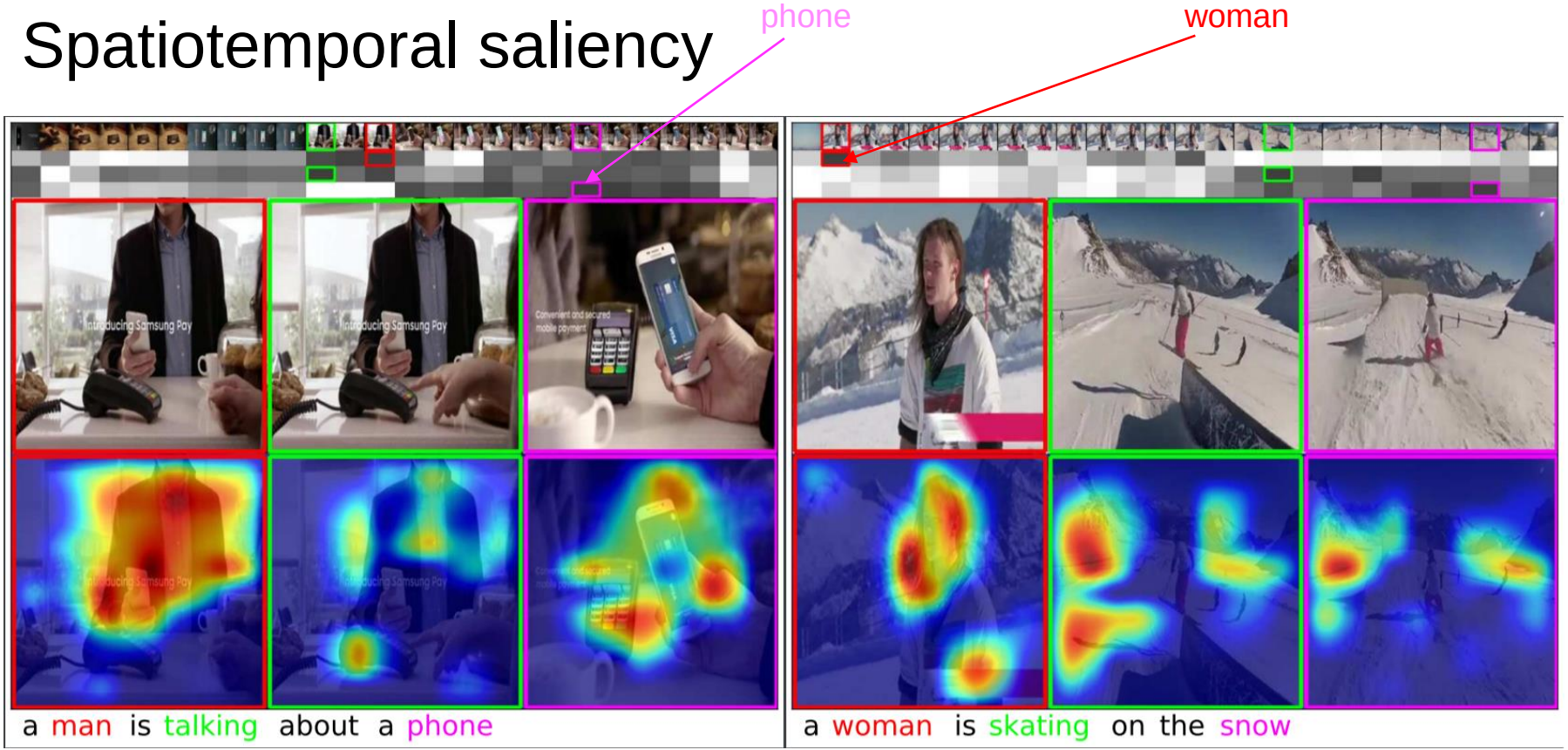


Image captioning with the same architecture

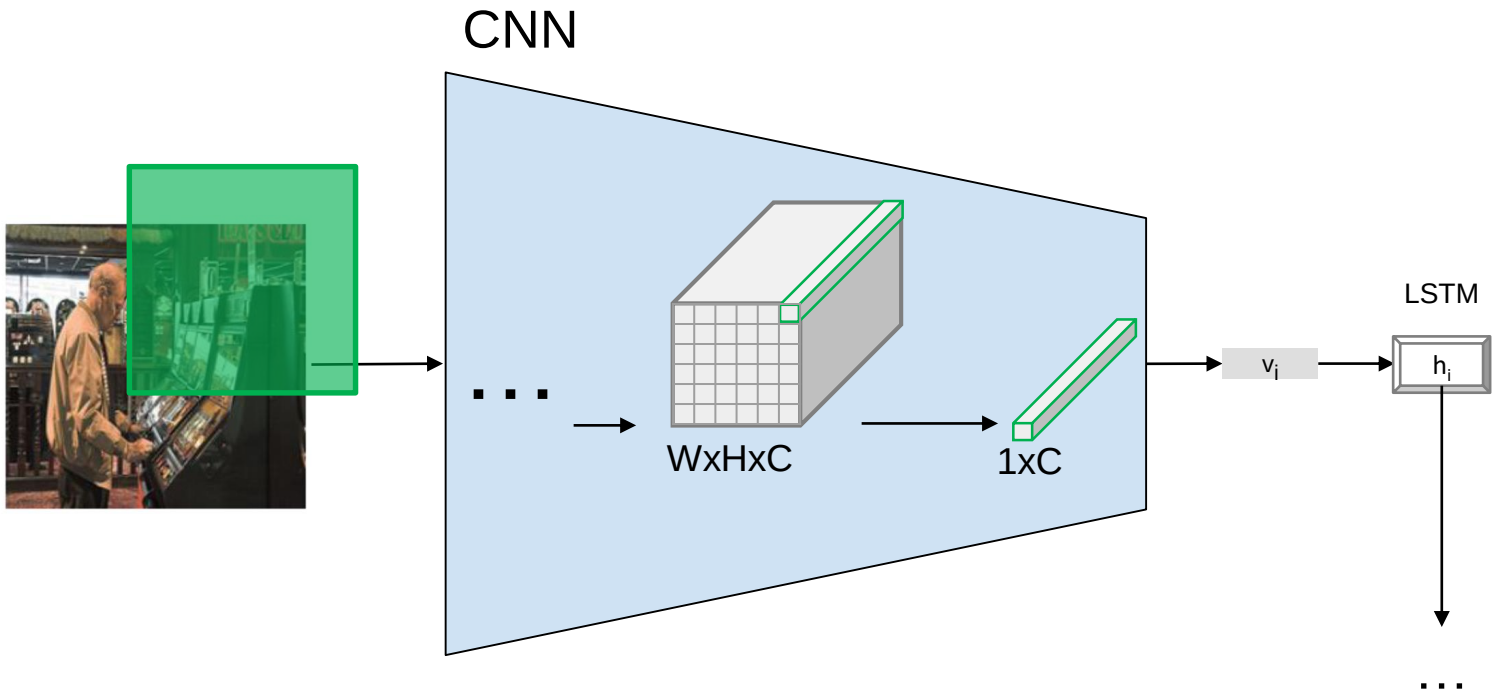
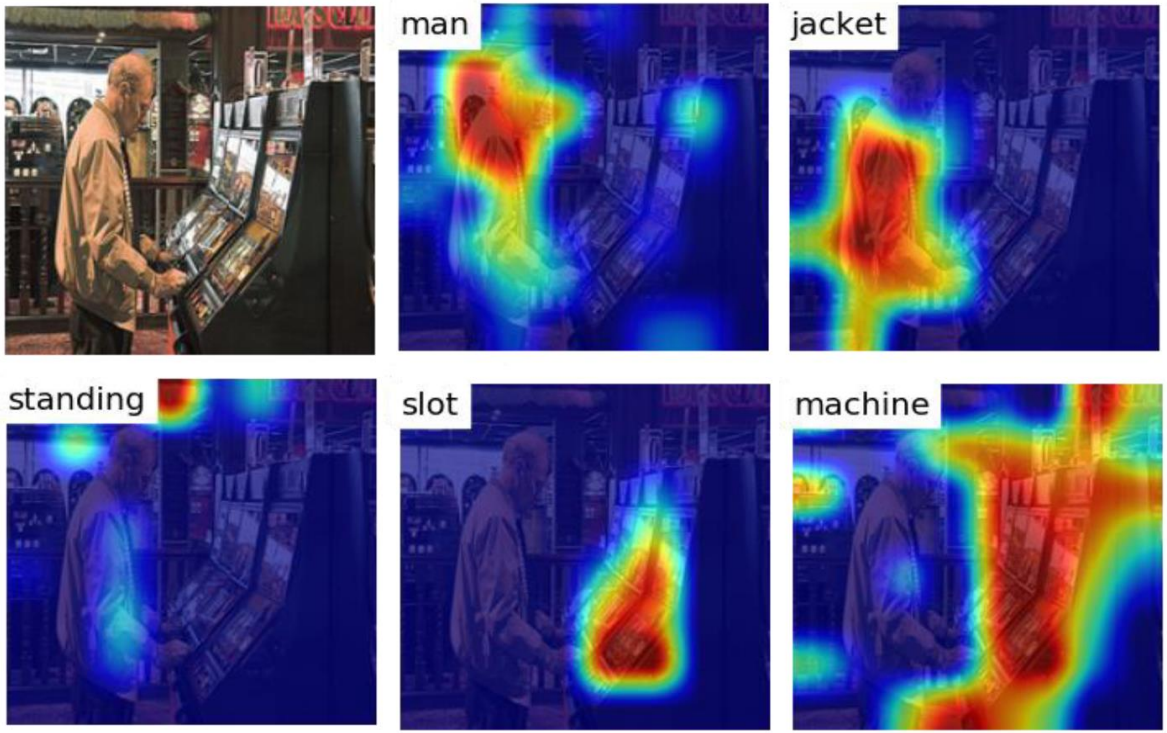
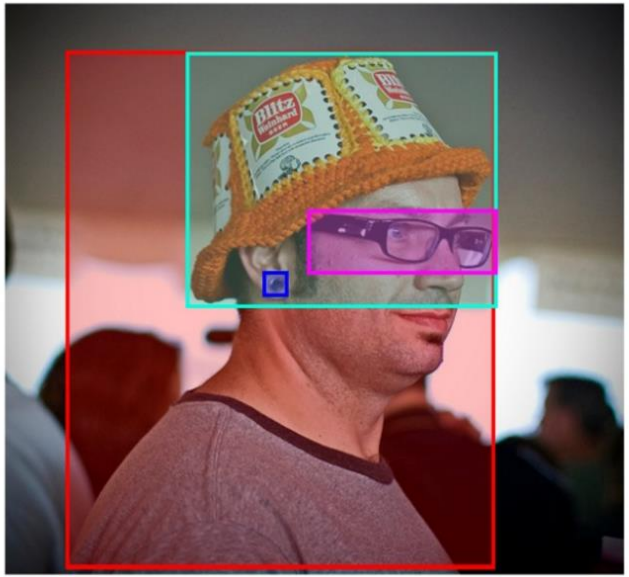


Image captioning with the same architecture

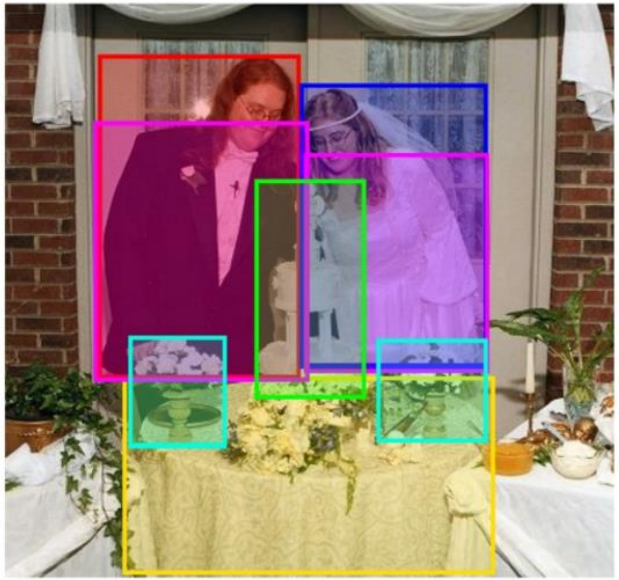
Input query: A **man** in a **jacket** is **standing** at the **slot machine**



Flickr30kEntities



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

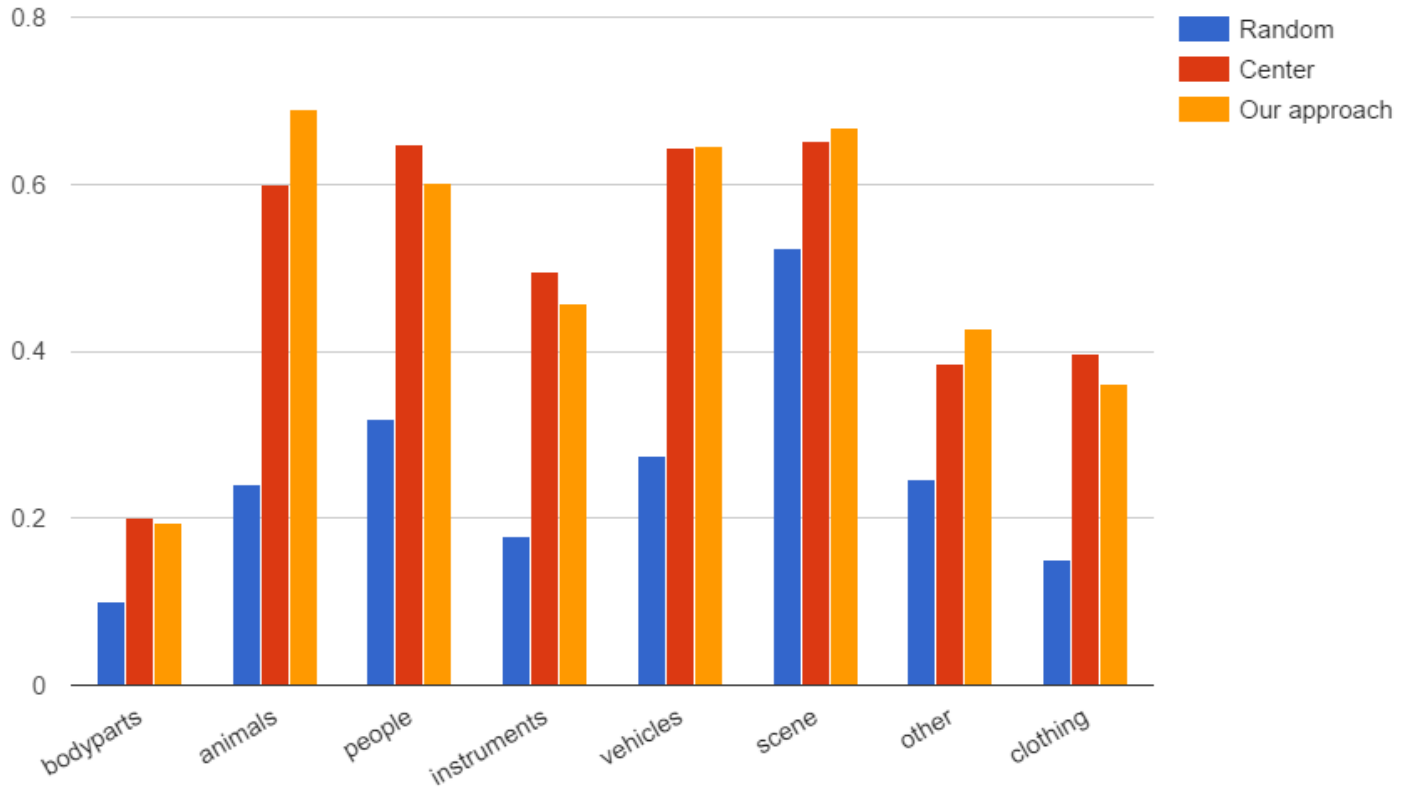


A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

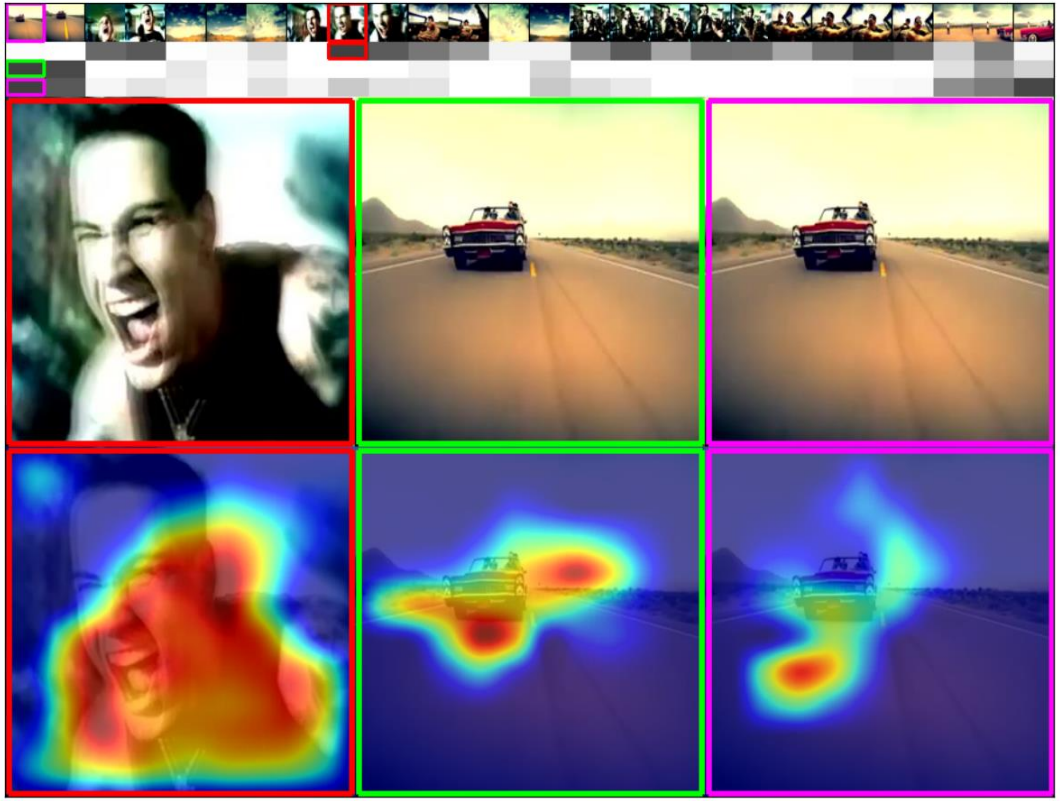
Plummer et al., ICCV 2015

Flickr30kEntities

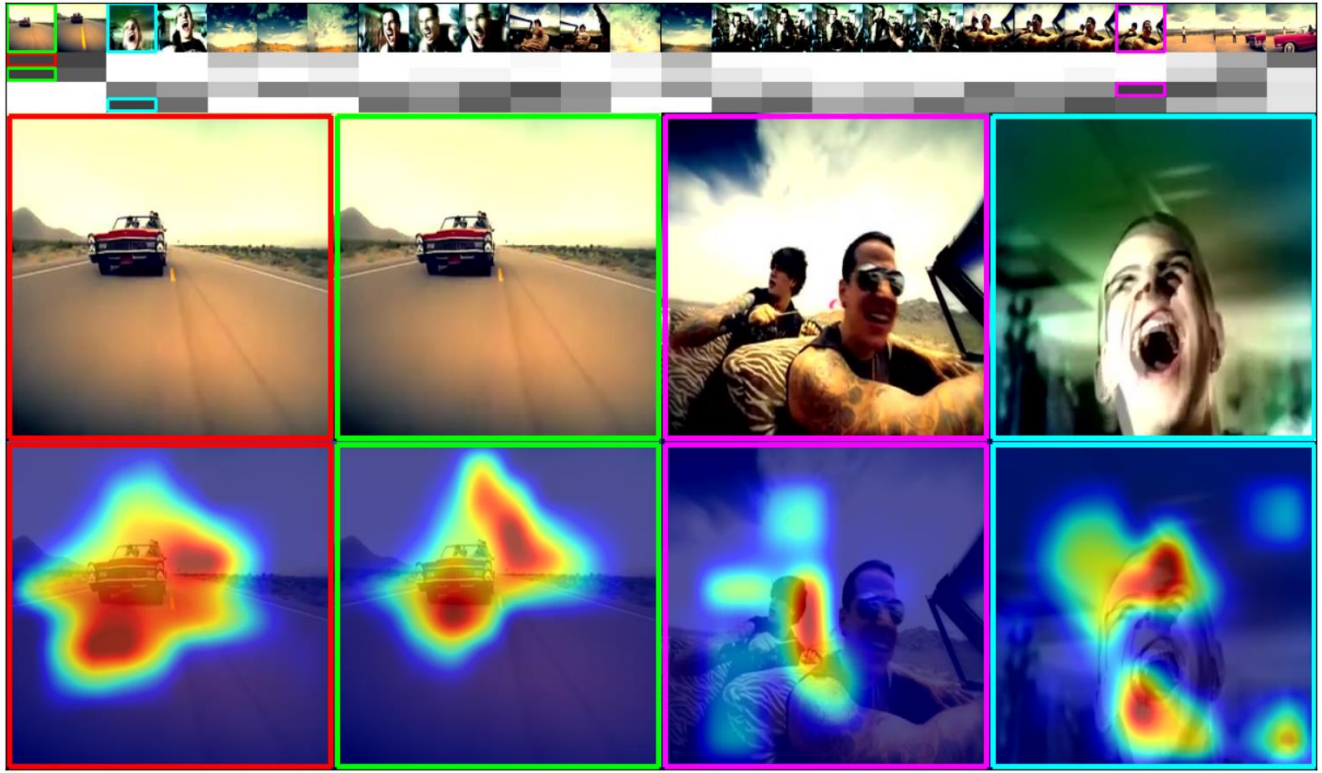
Flickr30kEntities



Video summarization: predicted sentence



Video summarization: arbitrary query



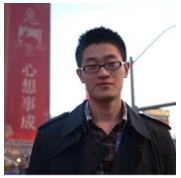
Video summarization: arbitrary query

Video summarization: arbitrary query

Thanks

Abir Das
Boston University

Marcus Rohrbach
UC Berkeley



Jianming Zhang
Adobe Research

Subhashini Venugopalan
UT Austin

Lisa Hendricks
UC Berkeley

Kate Saenko
Boston University