Why Did You Say That? Explaining and Diversifying Captioning Models

Kate Saenko

BOSTON UNIVERSITY

VQA Workshop, CVPR, July 26, 2017

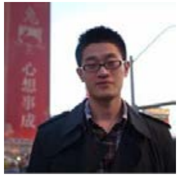
Explaining:

Top-down saliency guided by captions

http://ai.bu.edu/caption-guided-saliency/

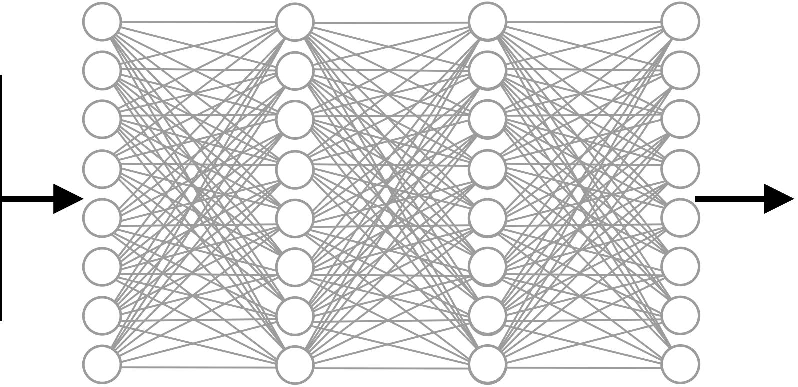

Vasili
Ramanishka
Boston University


Abir
Das
Boston University


Jianming
Zhang
Adobe Research


Kate
Saenko
Boston University

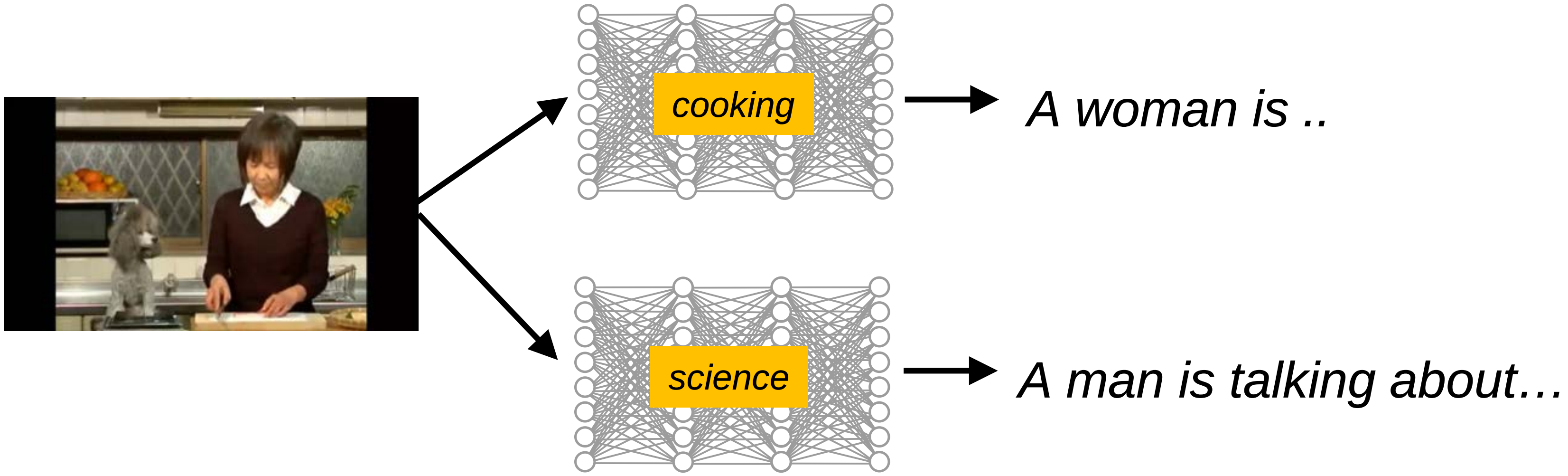
Captioning

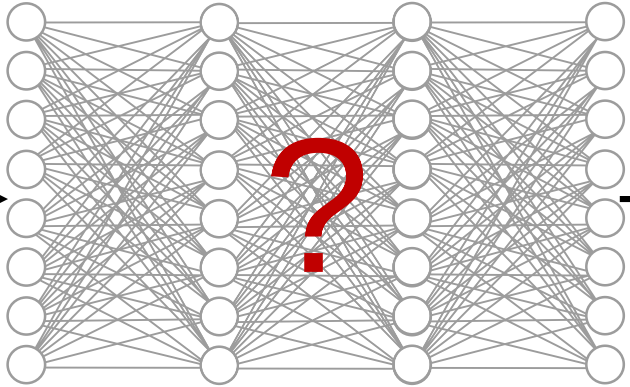


A woman is cutting a piece of meat

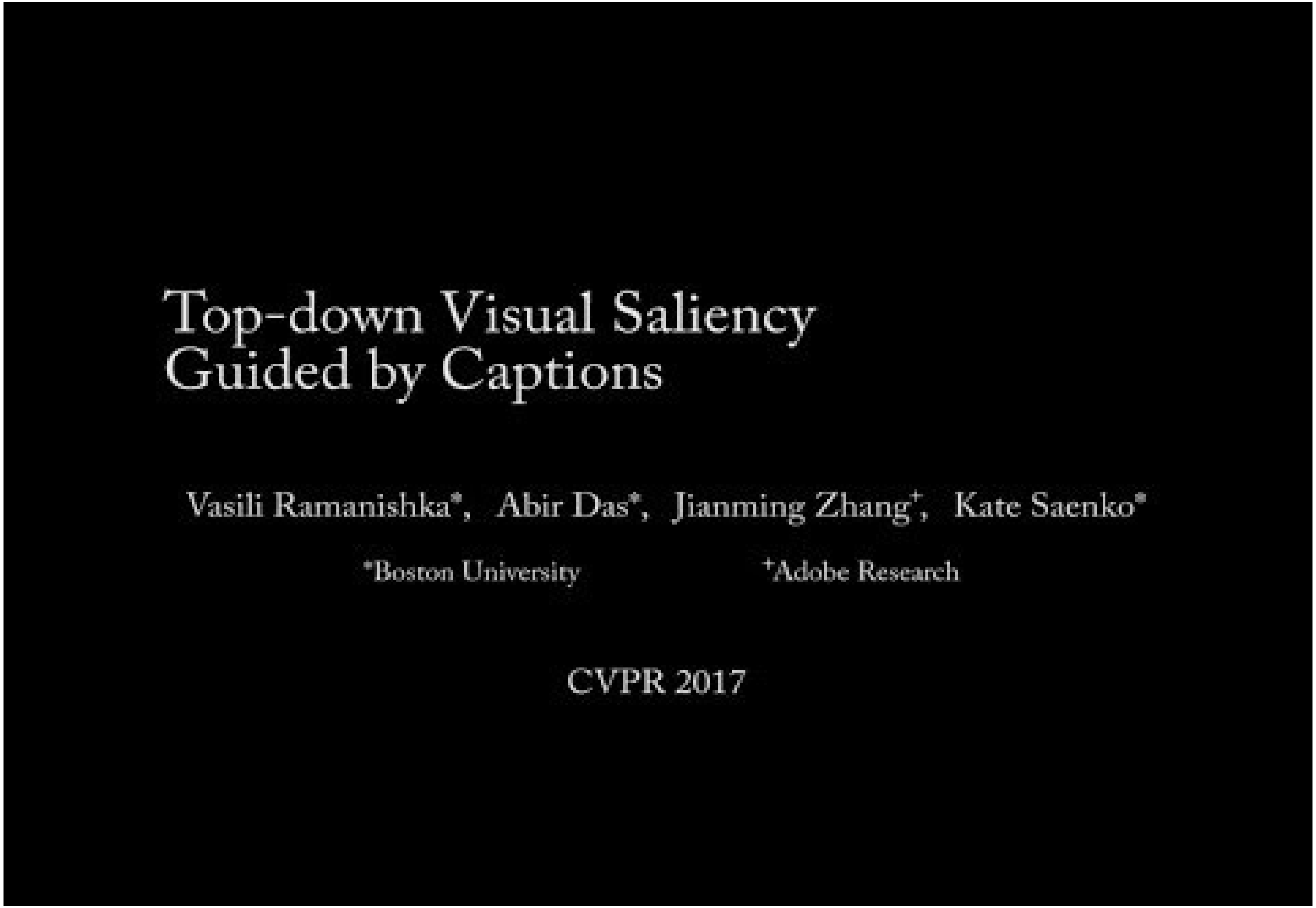
Why did the network say that?

Captioning





A woman is cutting a piece of meat


Top-down Visual Saliency
Guided by Captions
Vasili Ramanishka*, Abir Das*, Jianming Zhang+, Kate Saenko*
*Boston University
+Adobe Research
CVPR 2017

Explaining the network's captions

Predicted sentence: A woman is cutting a piece of meat

can the network localize objects?

Related: Attention layers

“Attention Layers”: Sequentially process regions in a single image.
Objective: Model learns “where to look” next.

Image Captioning

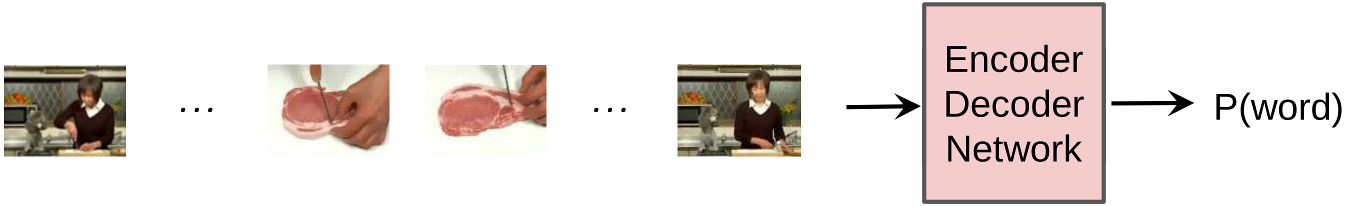
girl teddy bear

Show, Attend and Tell
[Xu et al. ICML’15]

- *soft attention* adds special attention layer
- Only spatial or only temporal

- Hard to do spatio-temporal attention
- Can we get salient regions without adding such layers?

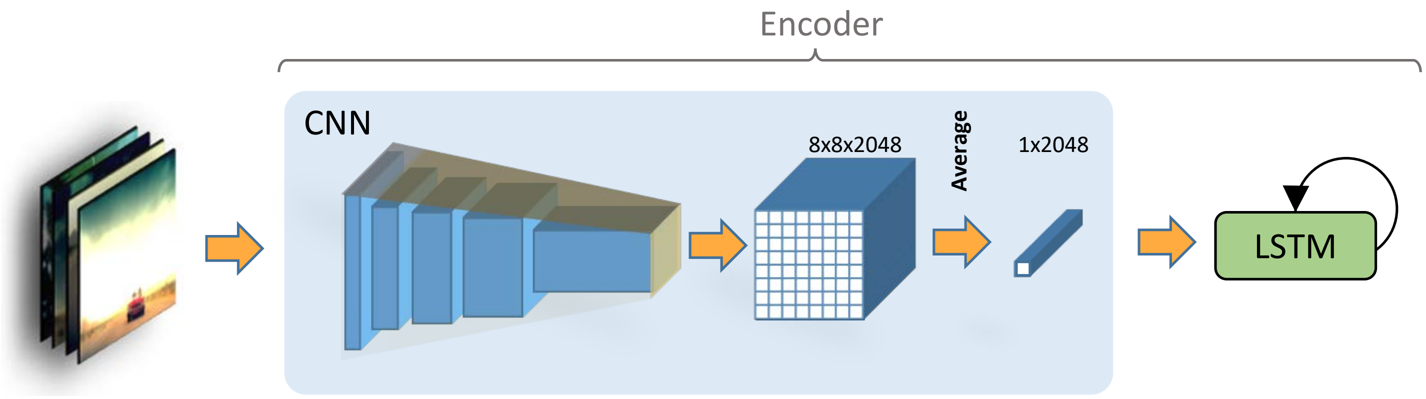
Key idea:
probe the network with small part of input



- *No need for* special attention layer
- Get spatio-temporal attention for free

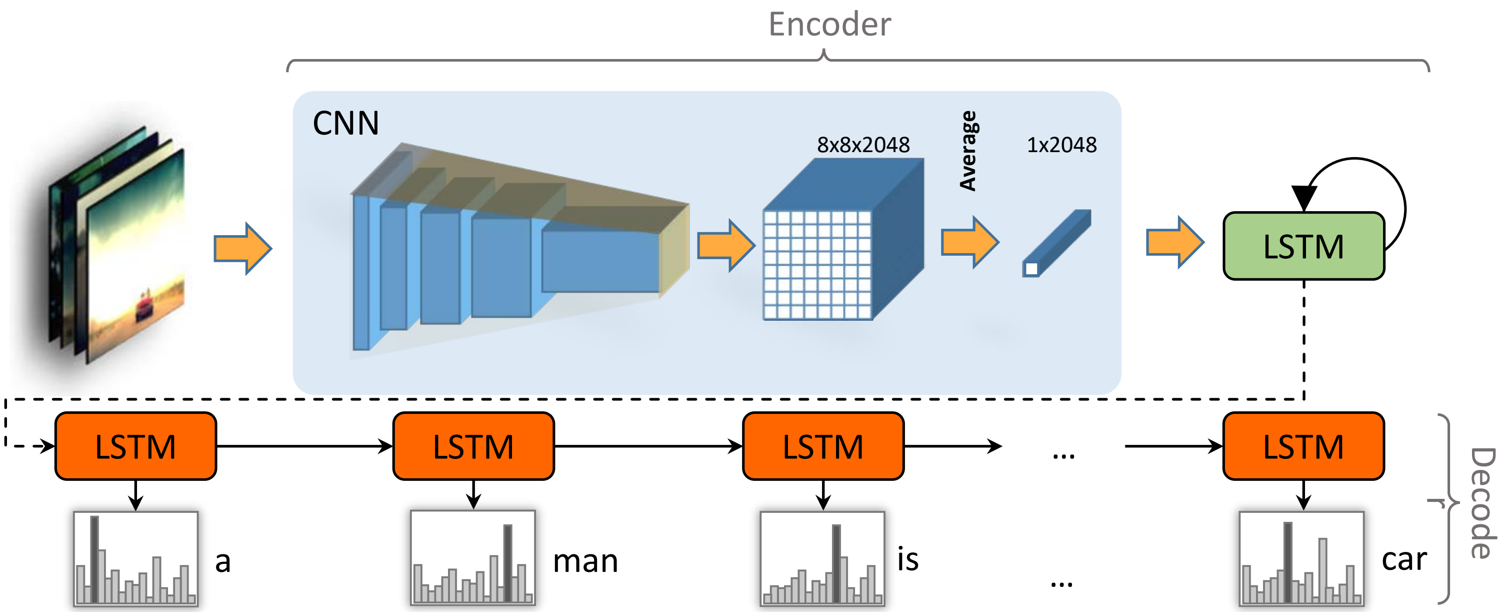
Encoder-decoder framework for video description

slide: Vasili Ramanishka



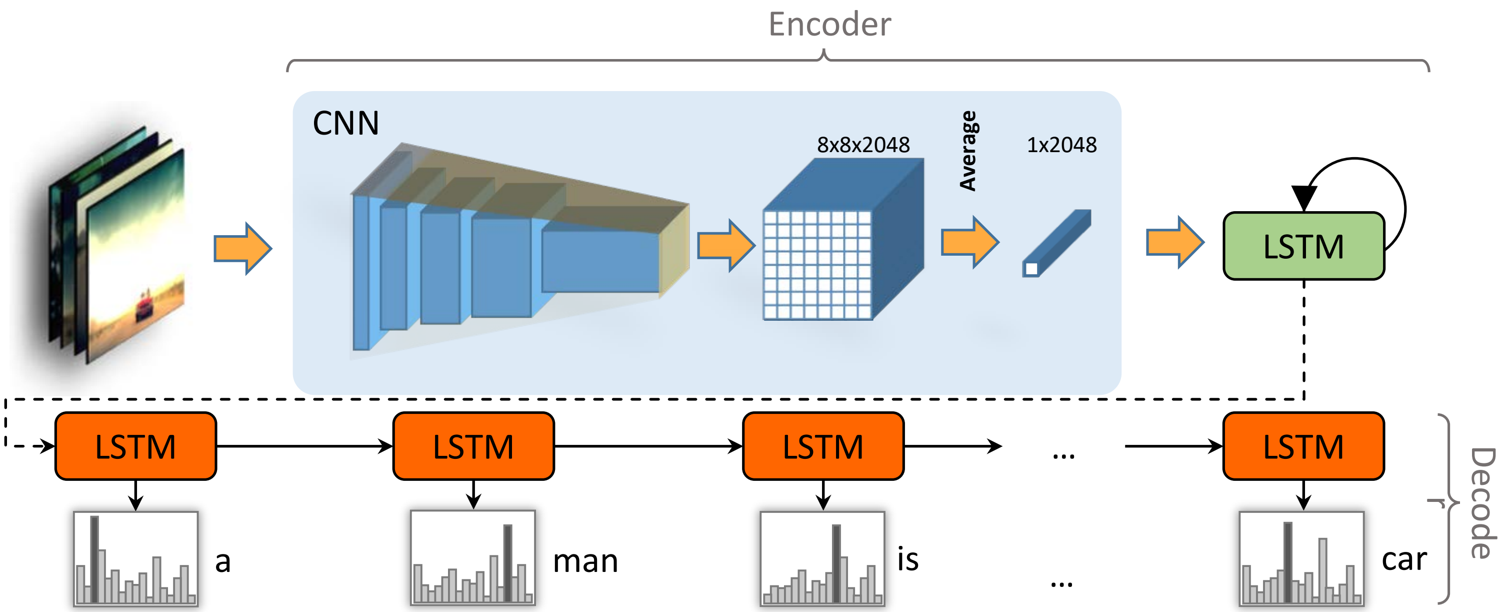
Encoder-decoder framework for video description

slide: Vasili Ramanishka

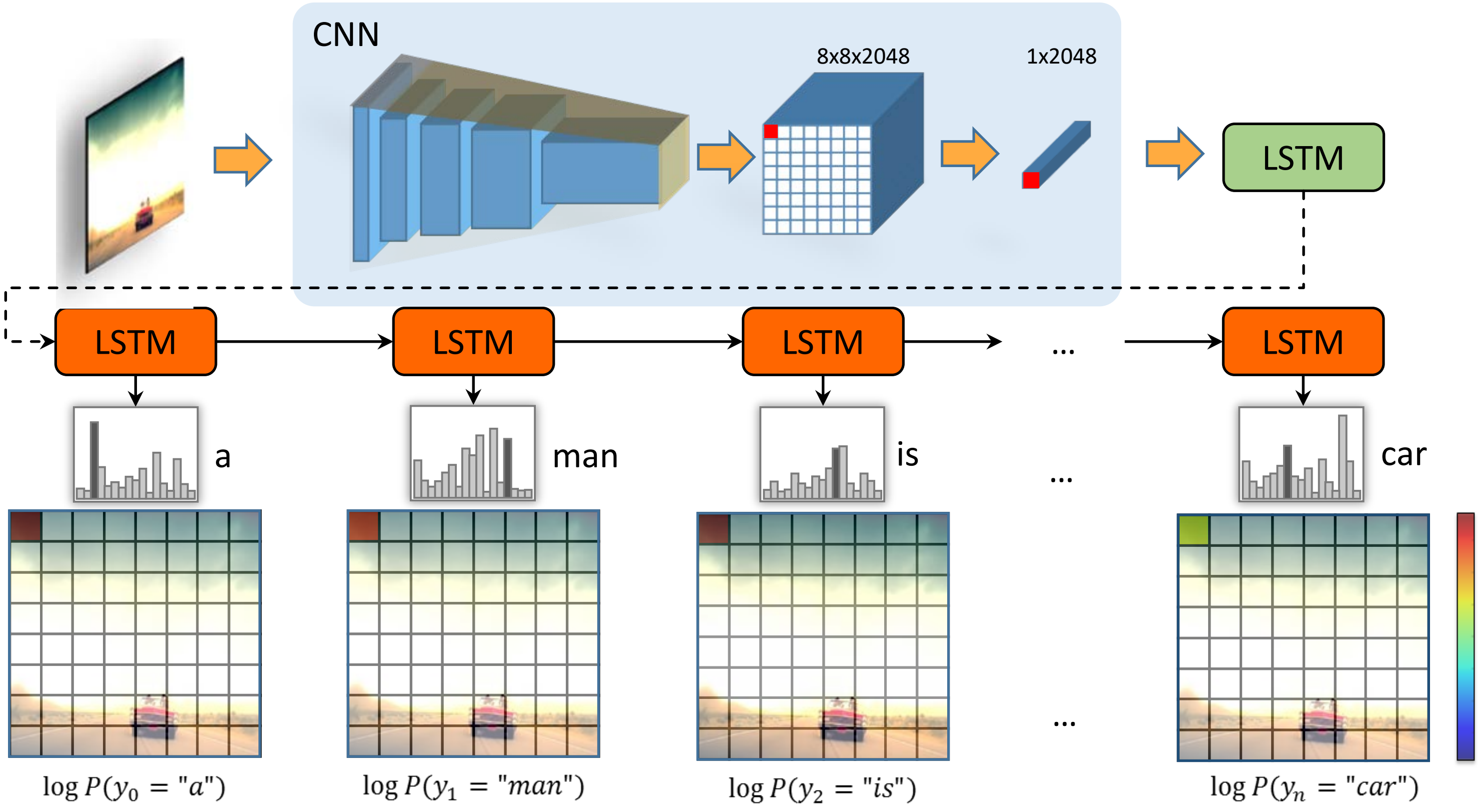


Encoder-decoder framework for video description

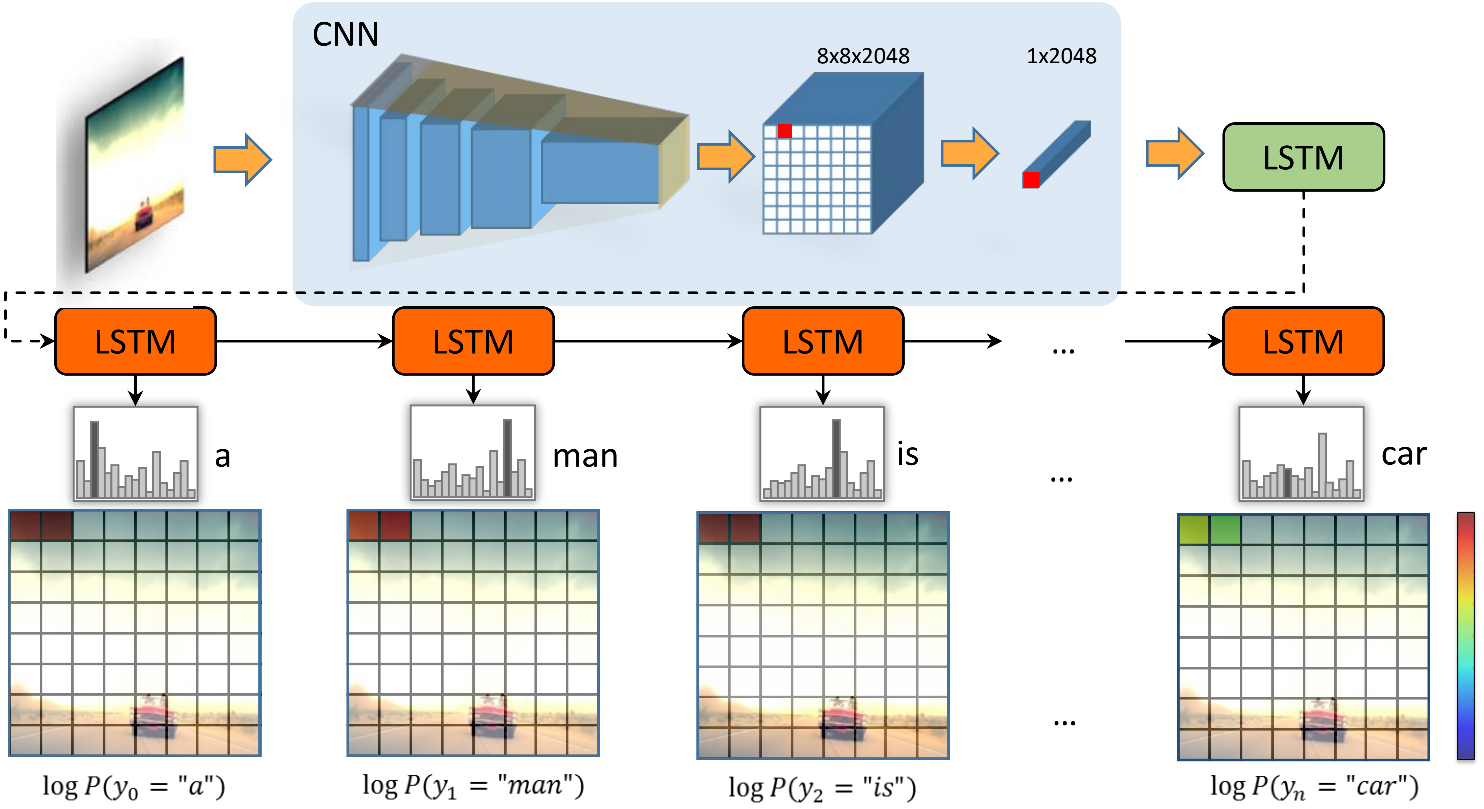
slide: Vasili Ramanishka



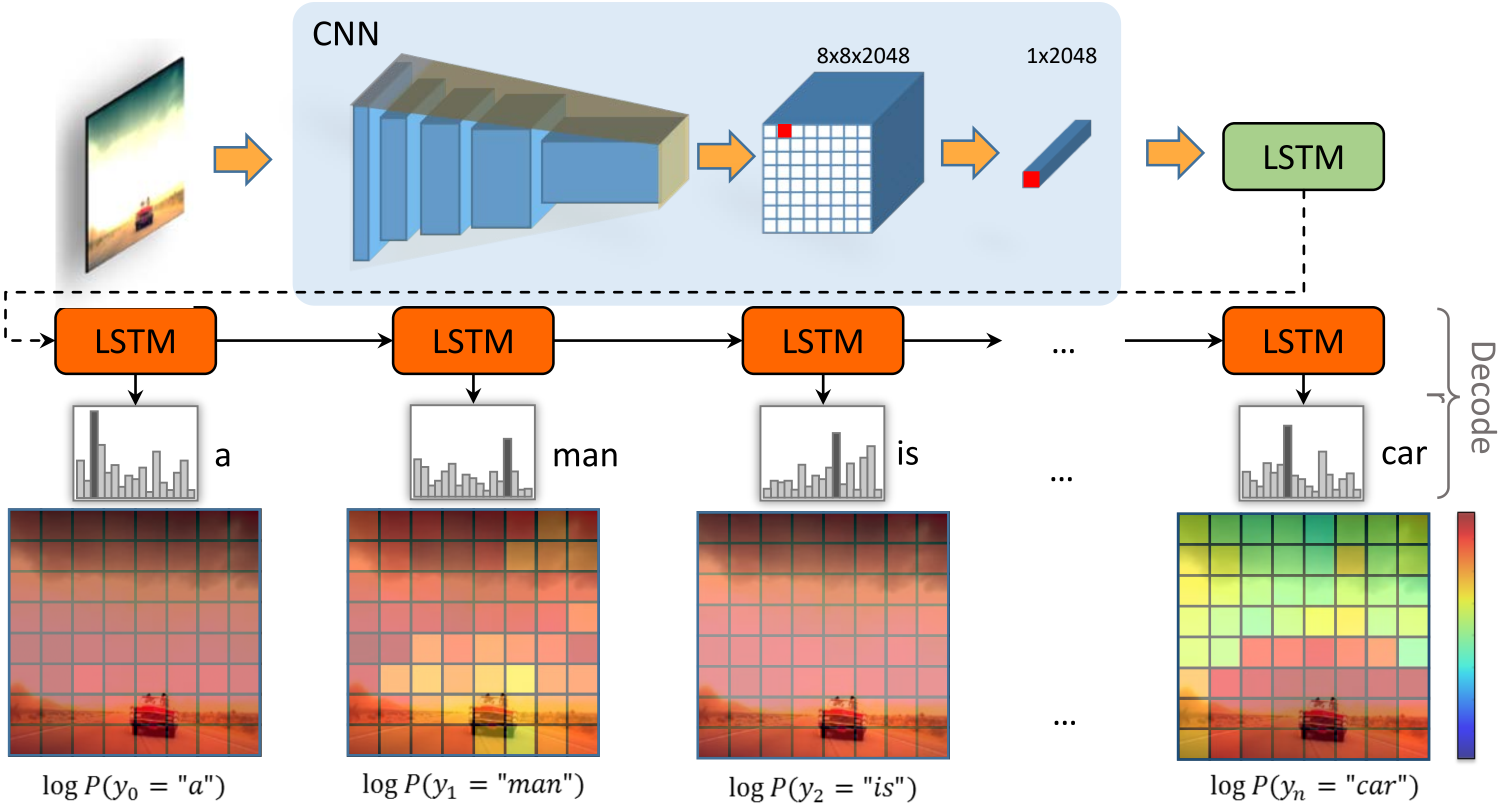
Saliency Estimation



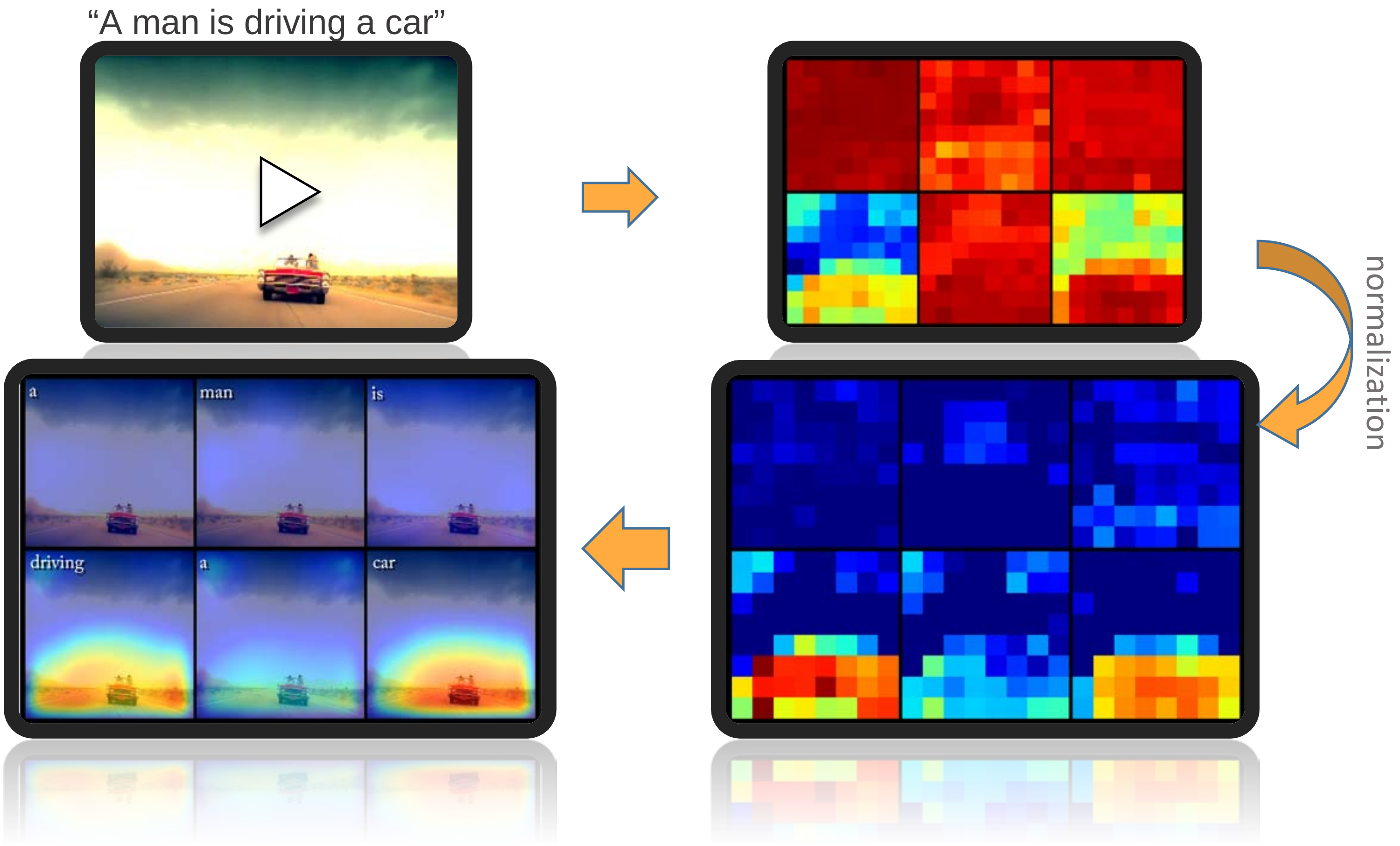
Saliency Estimation



Saliency Estimation

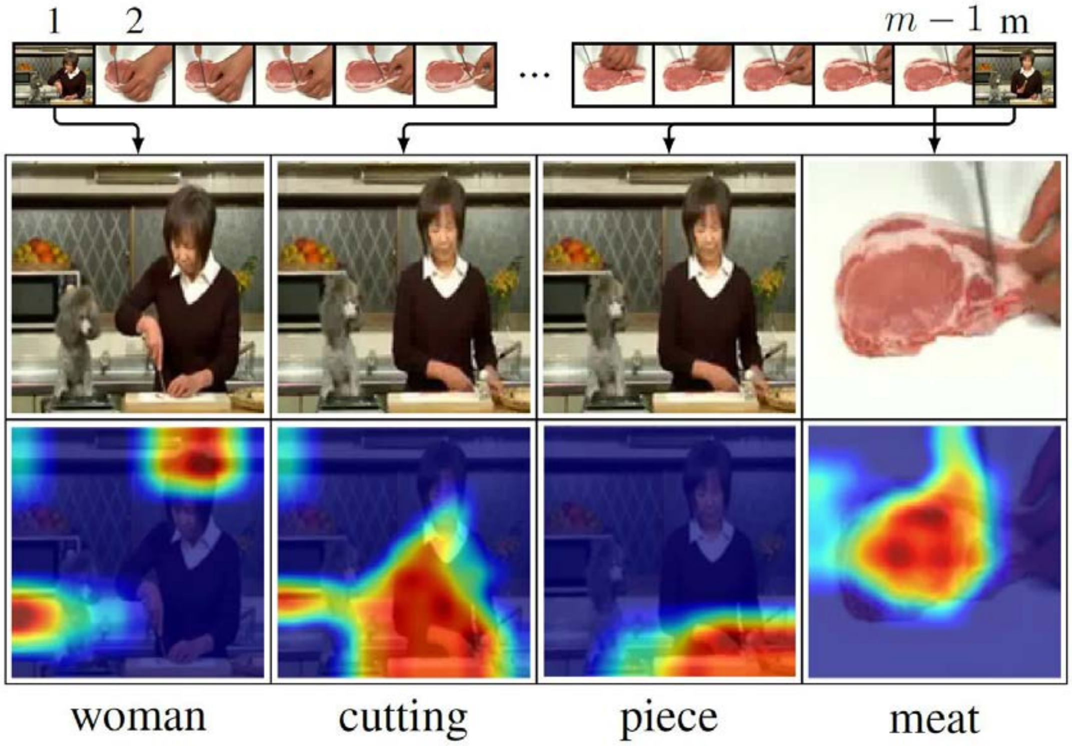


Saliency Estimation



Spatiotemporal saliency

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



Spatiotemporal saliency

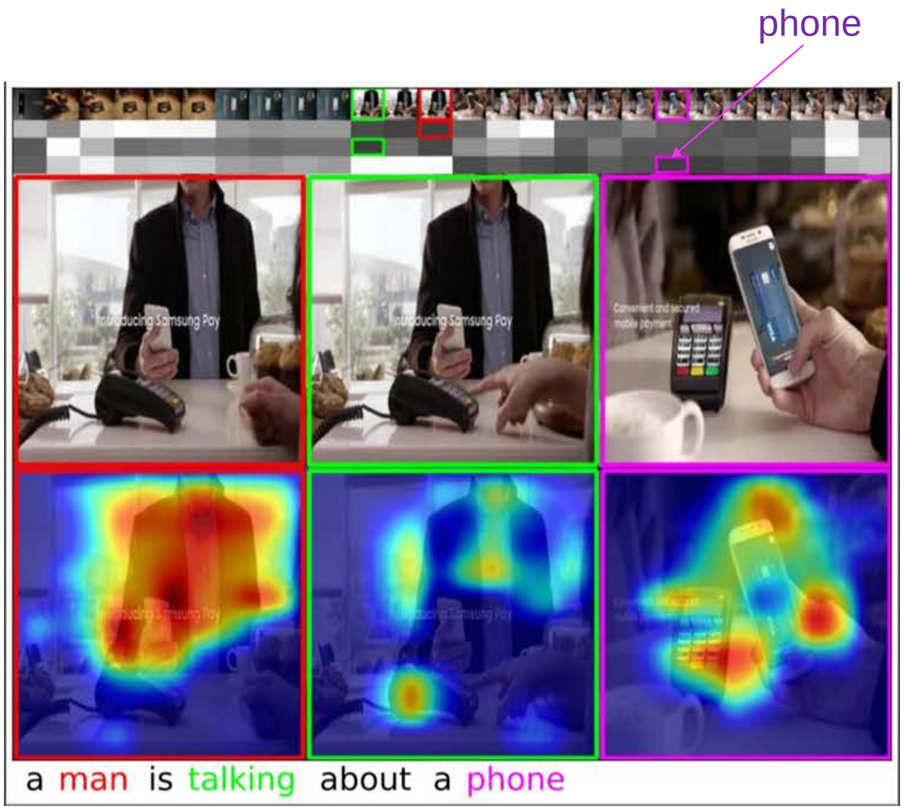


Image captioning with the same architecture

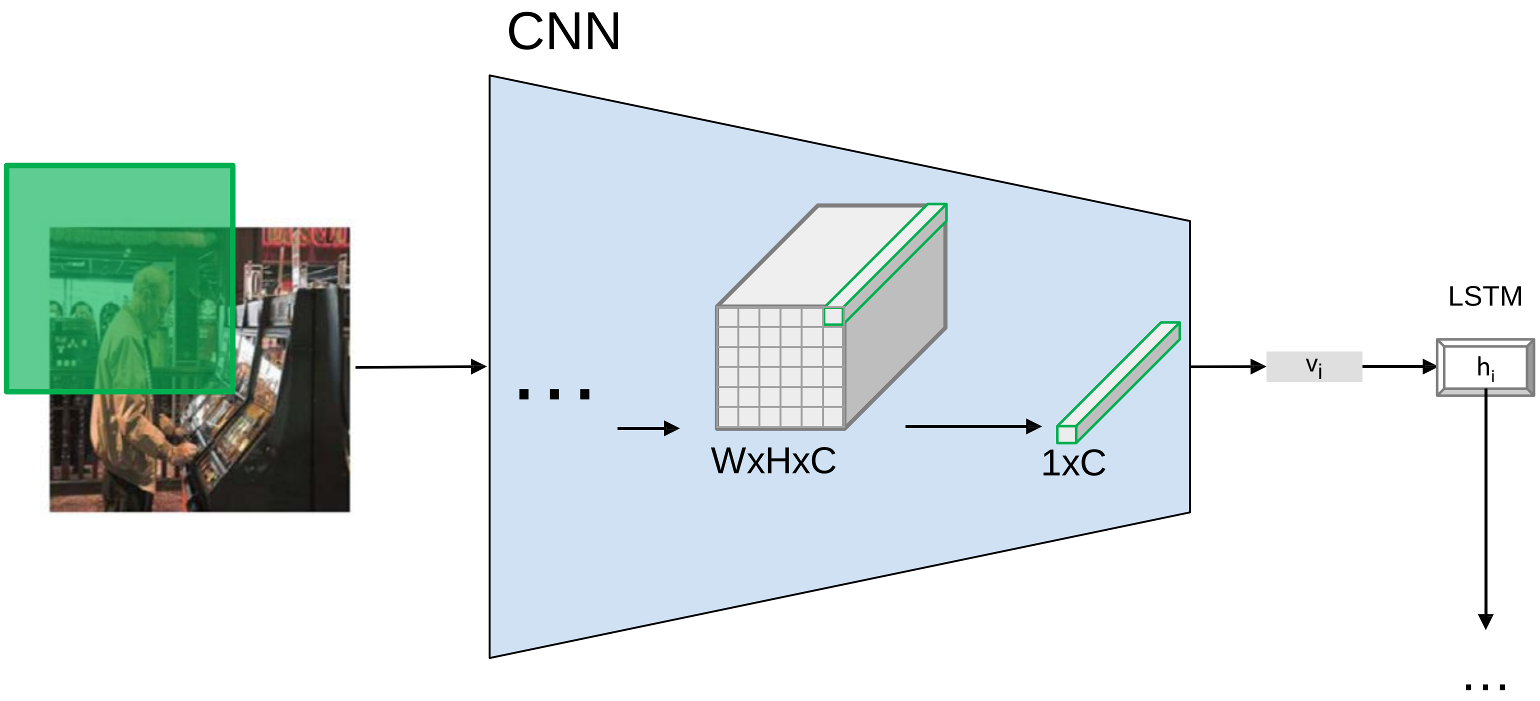
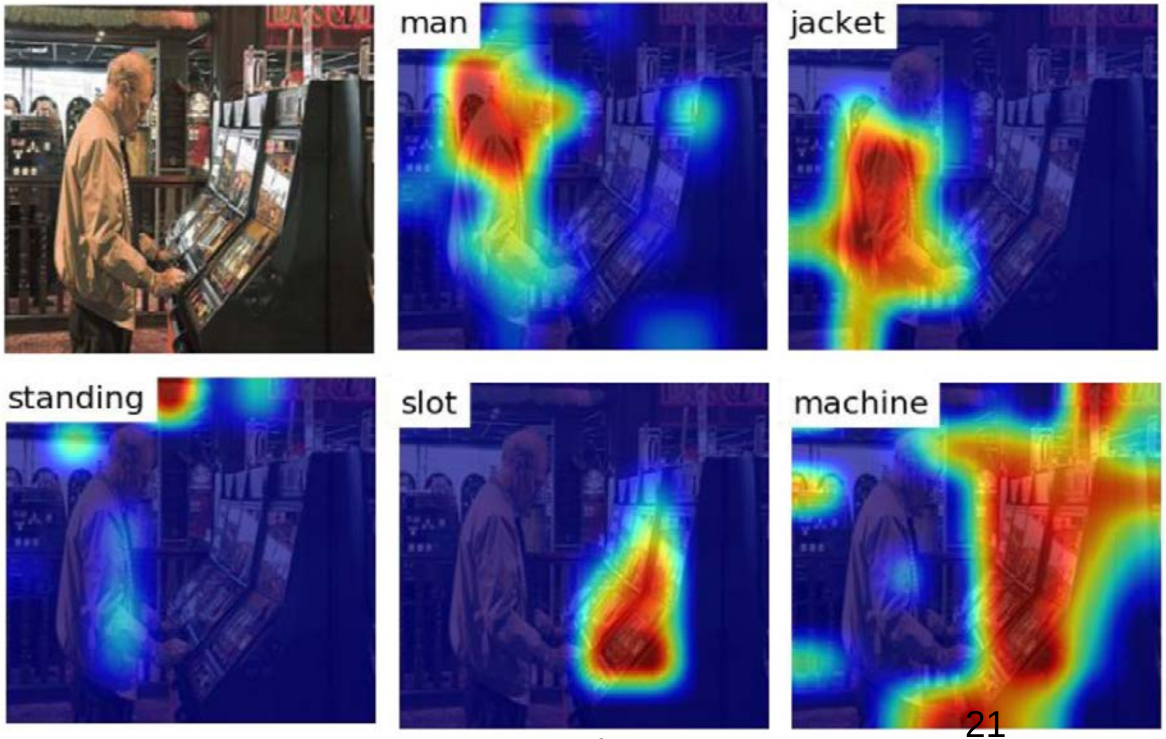


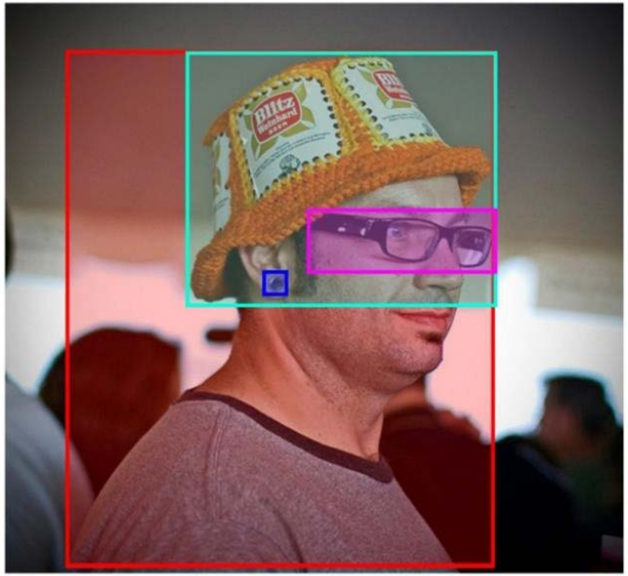
Image captioning with the same architecture

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

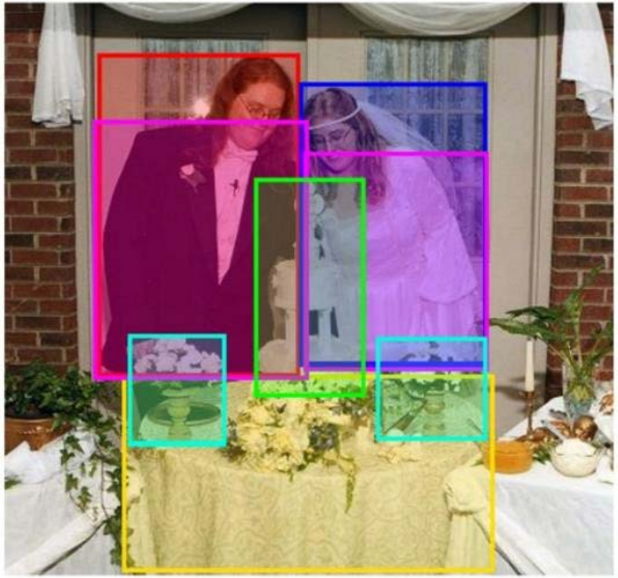


Kate Saenko

Flickr30kEntities



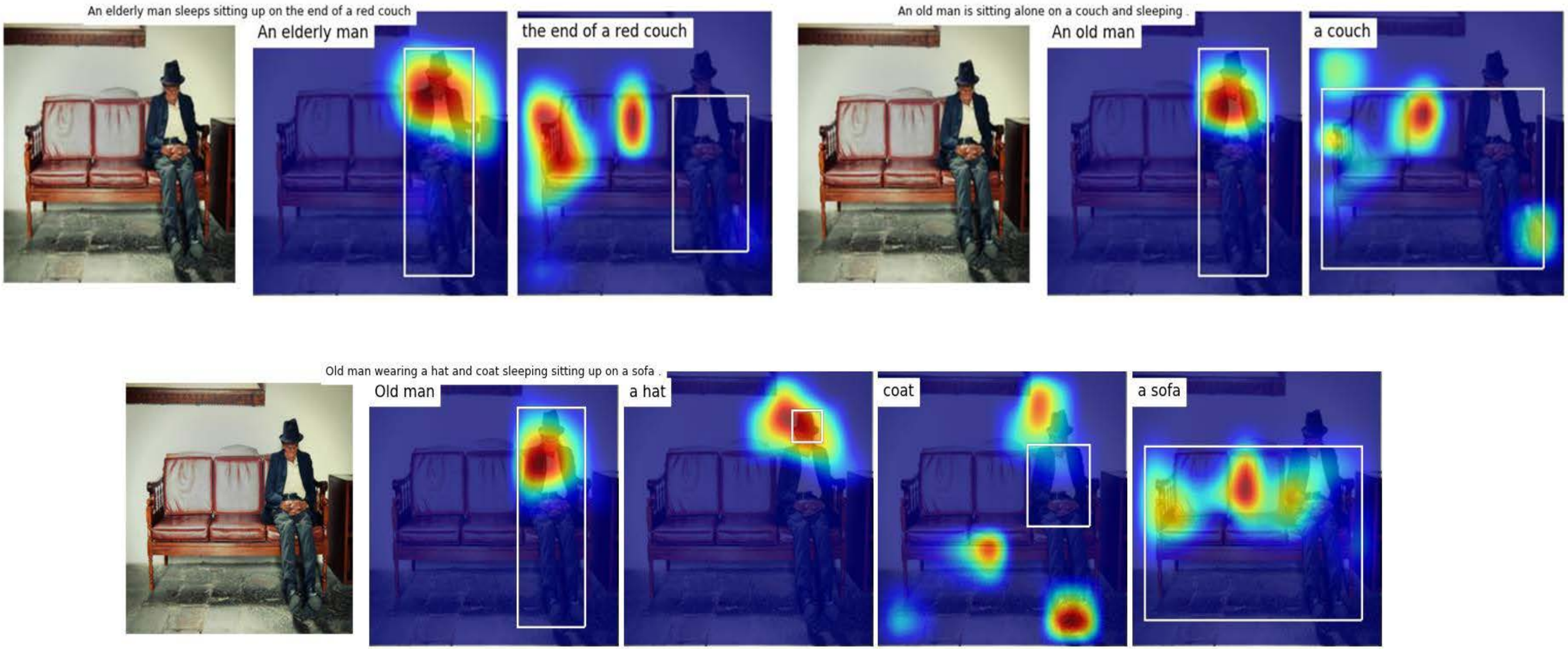
A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.



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

Pointing game in Flickr30kEntities



Comparison to Soft Attention on Flickr30kEntities

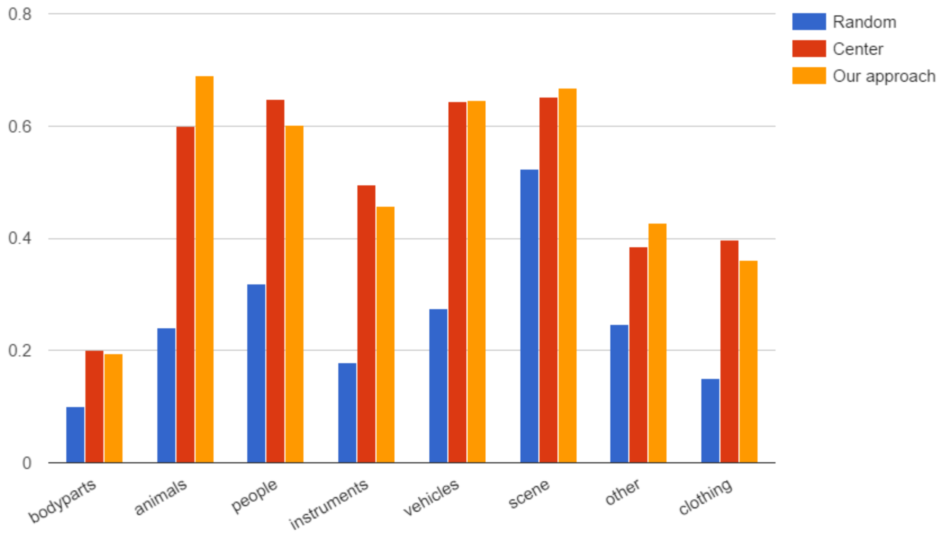
Attention correctness

	Avg per NP
Baseline [14]	0.321
SA [14]	0.387
SA-supervised [14]	0.433
Baseline*	0.325
Our model	**0.473**

Captioning performance

Model	Dataset	METEOR [9]
Soft-Attn [28]	MSVD	30.0
Our Model	MSVD	31.0
Soft-Attn [12]	MSR-VTT	25.4
Our Model	MSR-VTT	25.9
Soft-Attn [27]	Flickr30k	18.5
Our Model	Flickr30k	18.3

Pointing game accuracy



[14] C. Liu, J. Mao, F. Sha, and A. L. Yuille. Attention correctness in neural image captioning, 2016, implementation of K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML 2015

Video summarization: predicted sentence

Video summarization: arbitrary query

Diversifying:

Captioning Images with Diverse Objects

Subhashini Venugopalan

Lisa Anne Hendricks

Marcus Rohrbach

Raymond Mooney

Kate Saenko

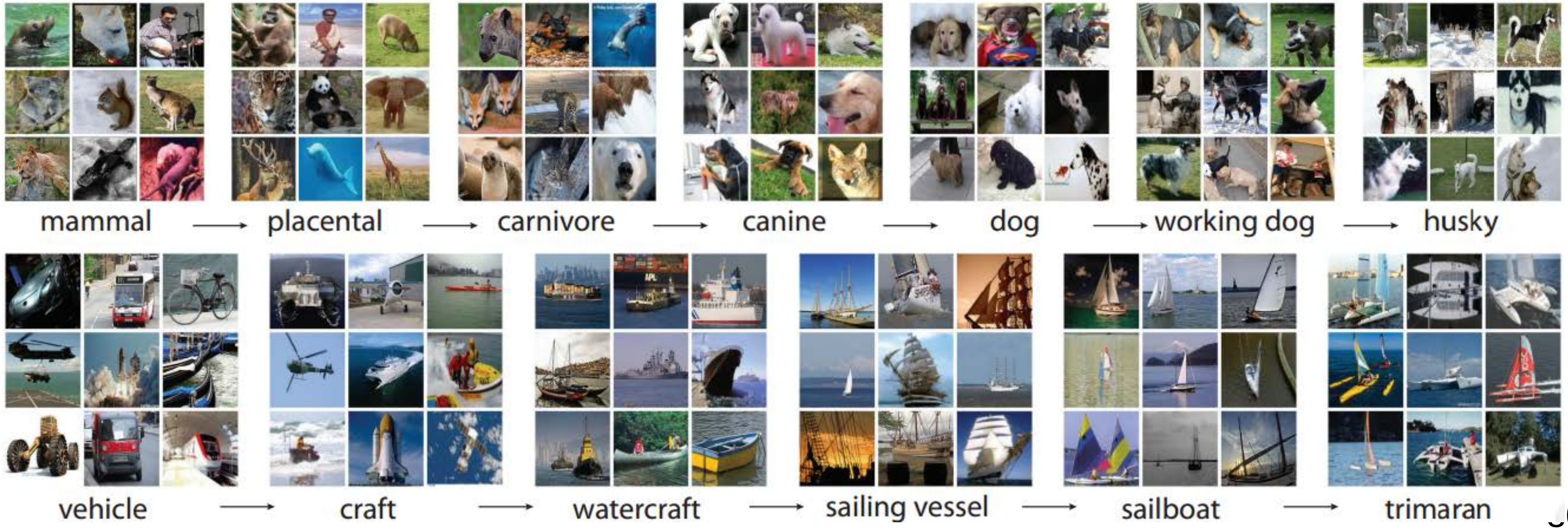
Trevor Darrell

UT Austin UC Berkeley Boston Univ.

Object Recognition

Can identify 1000's of categories of objects.

IMAGENET 14M images, 22K classes [Deng et al. CVPR'09]



Visual Description

Berkeley LRCN [Donahue et al. CVPR'15]:
A brown bear standing on top of a lush green field.

MSR CaptionBot [http://captionbot.ai/]:
A large brown bear walking through a forest.

MSCOCO
80 classes

Novel Object Captioner (NOC)

We present Novel Object Captioner which can compose descriptions of 100s of objects in context.

NOC (ours): Describe novel objects without paired image-caption data.

IMAGENET + MSCOCO +

→ An **okapi** standing in the middle of a field.

Visual Classifiers. IMAGENET

→ **okapi**

Existing captioners. IMAGENET init + train MSCOCO

→ A horse standing in the dirt.

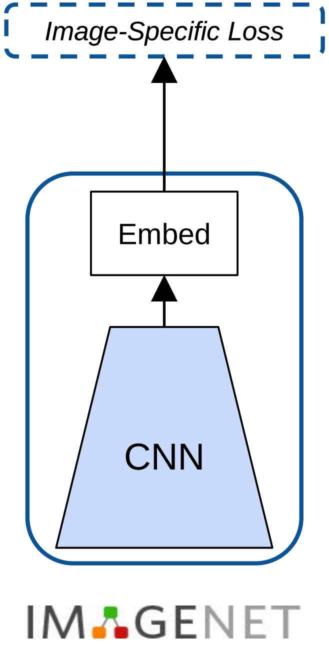
Insights

1. Need to recognize and describe objects outside of image-caption datasets.

→ okapi

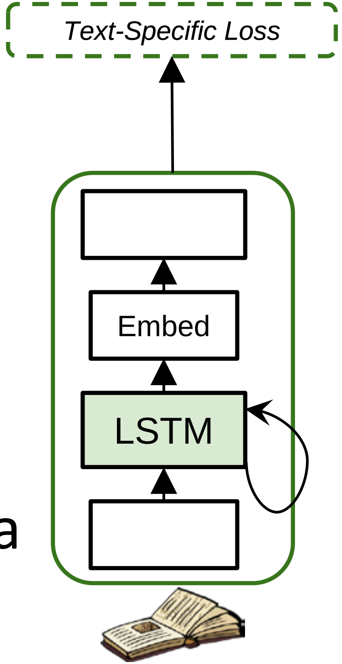
slide: Subhashini Venugopalan

Insight 1: Train effectively on external sources



Visual features from unpaired image data

Language model from unannotated text data



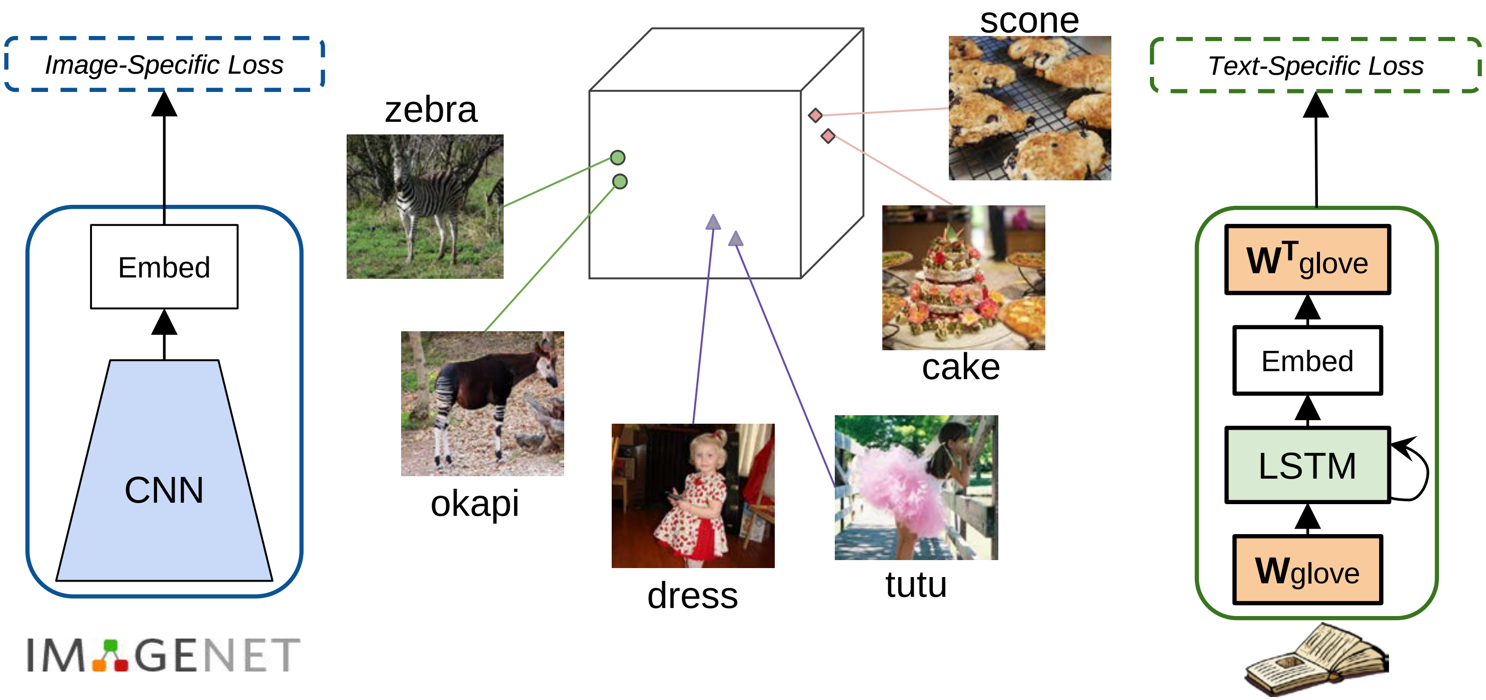
Insights

2. Describe unseen objects that are similar to objects seen in image-caption datasets.

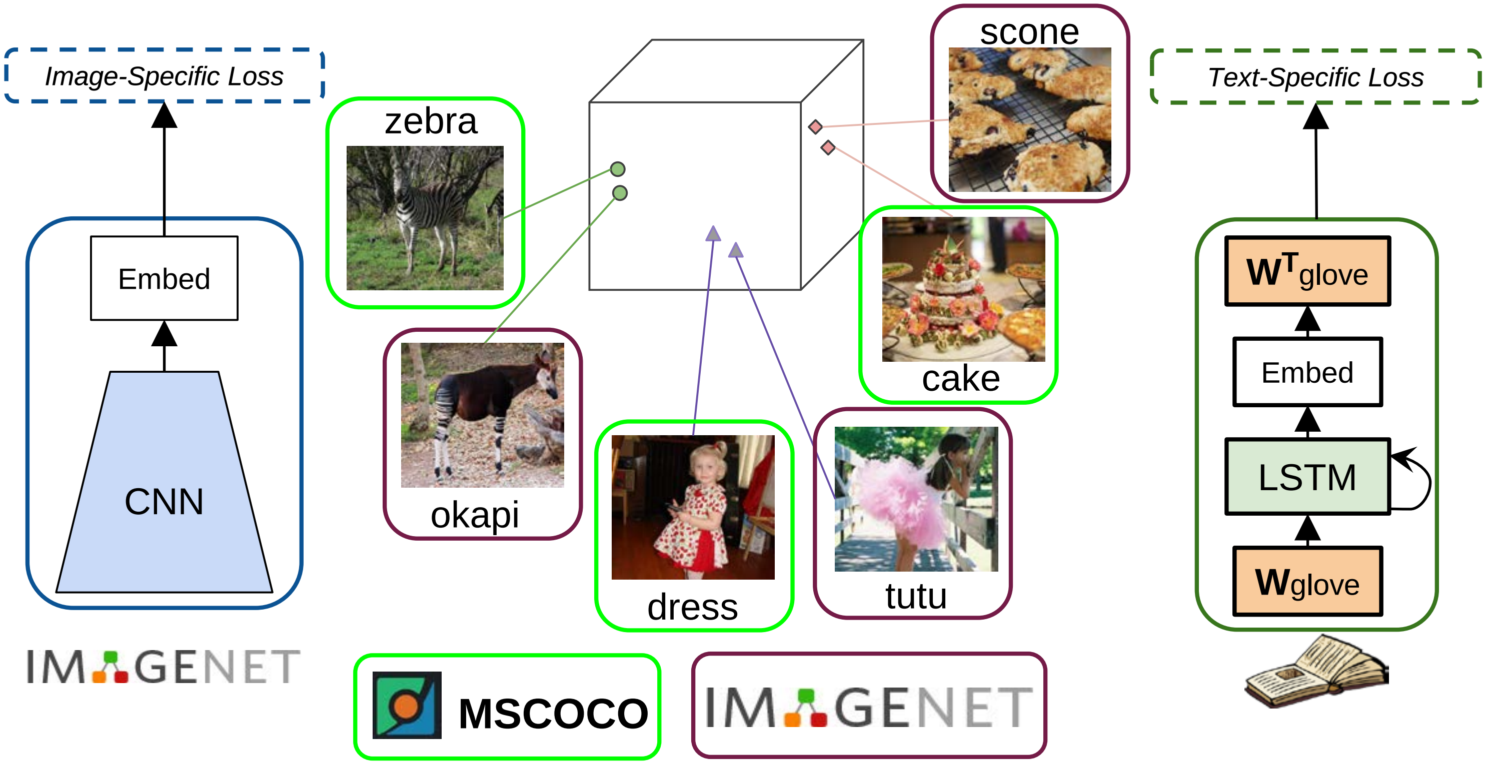
okapi

zebra

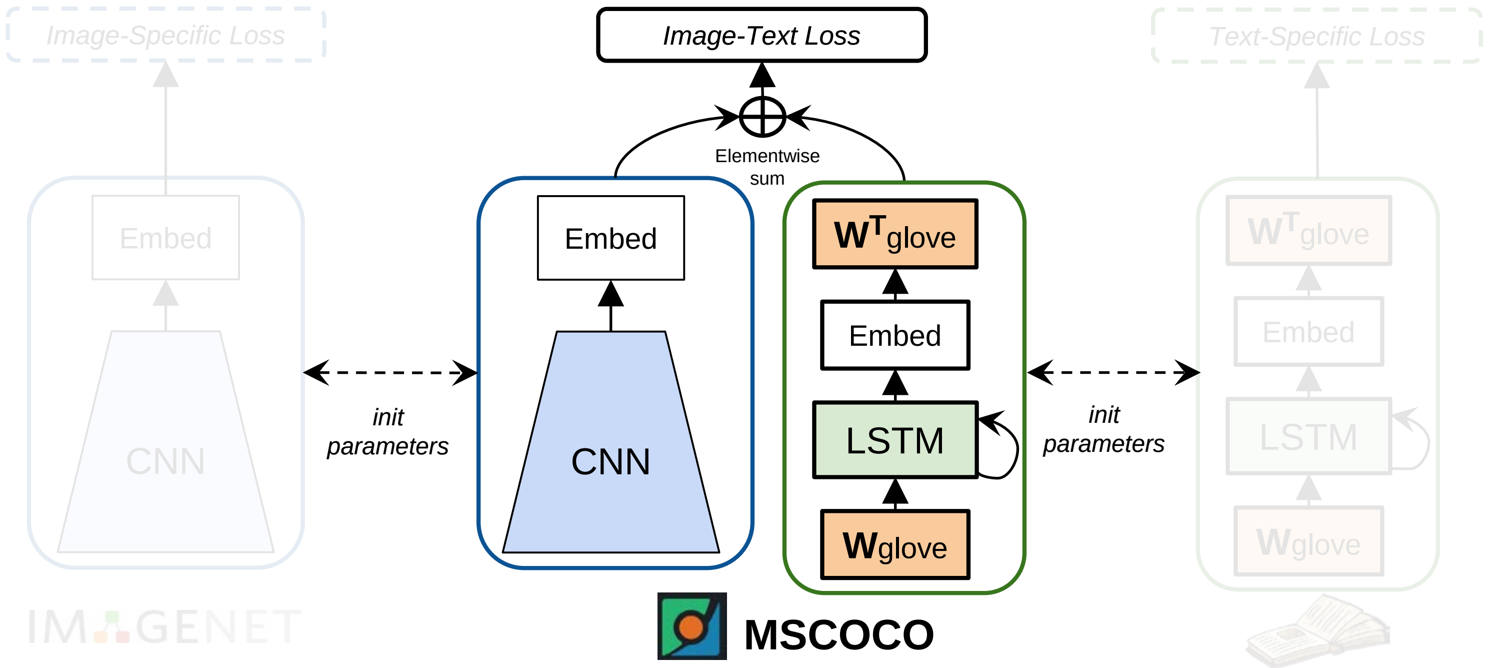
Insight 2: Capture semantic similarity of words



Insight 2: Capture semantic similarity of words



Combine to form a Caption Model



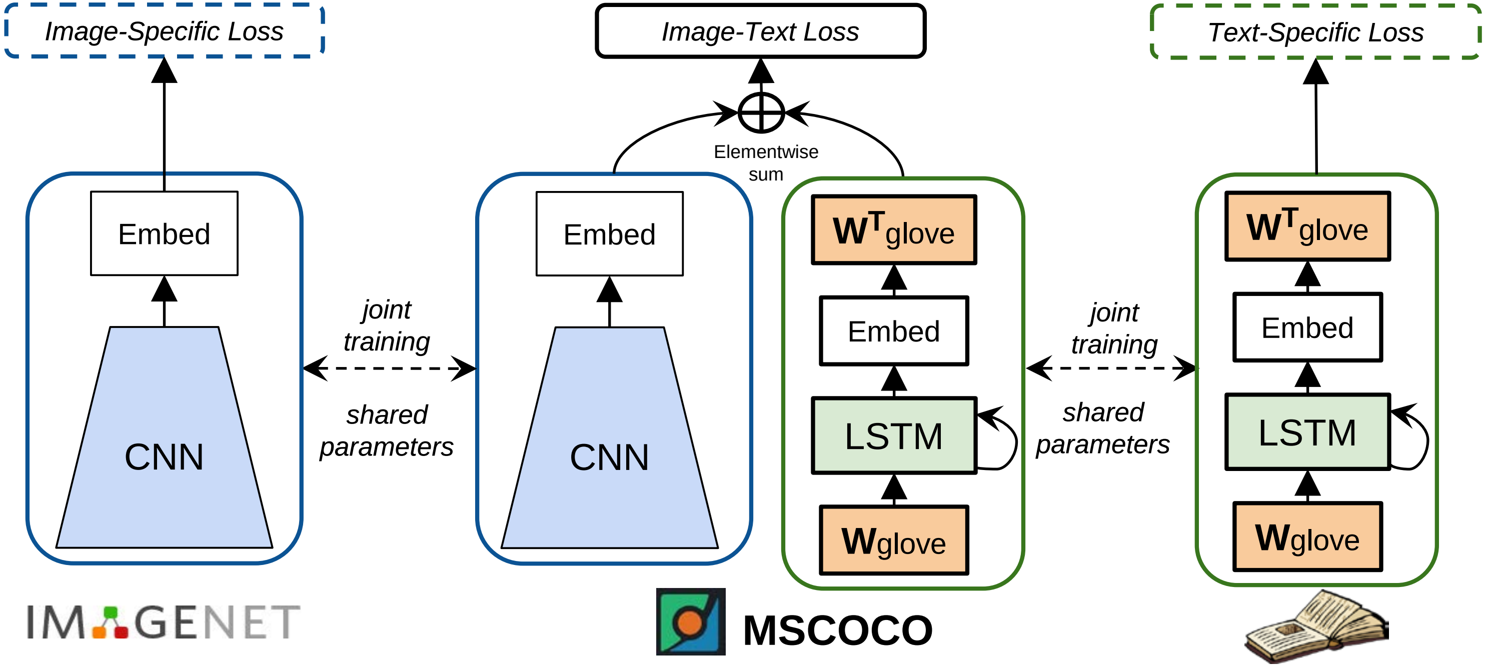
Not different from existing caption models. Problem: Forgetting.

Insights

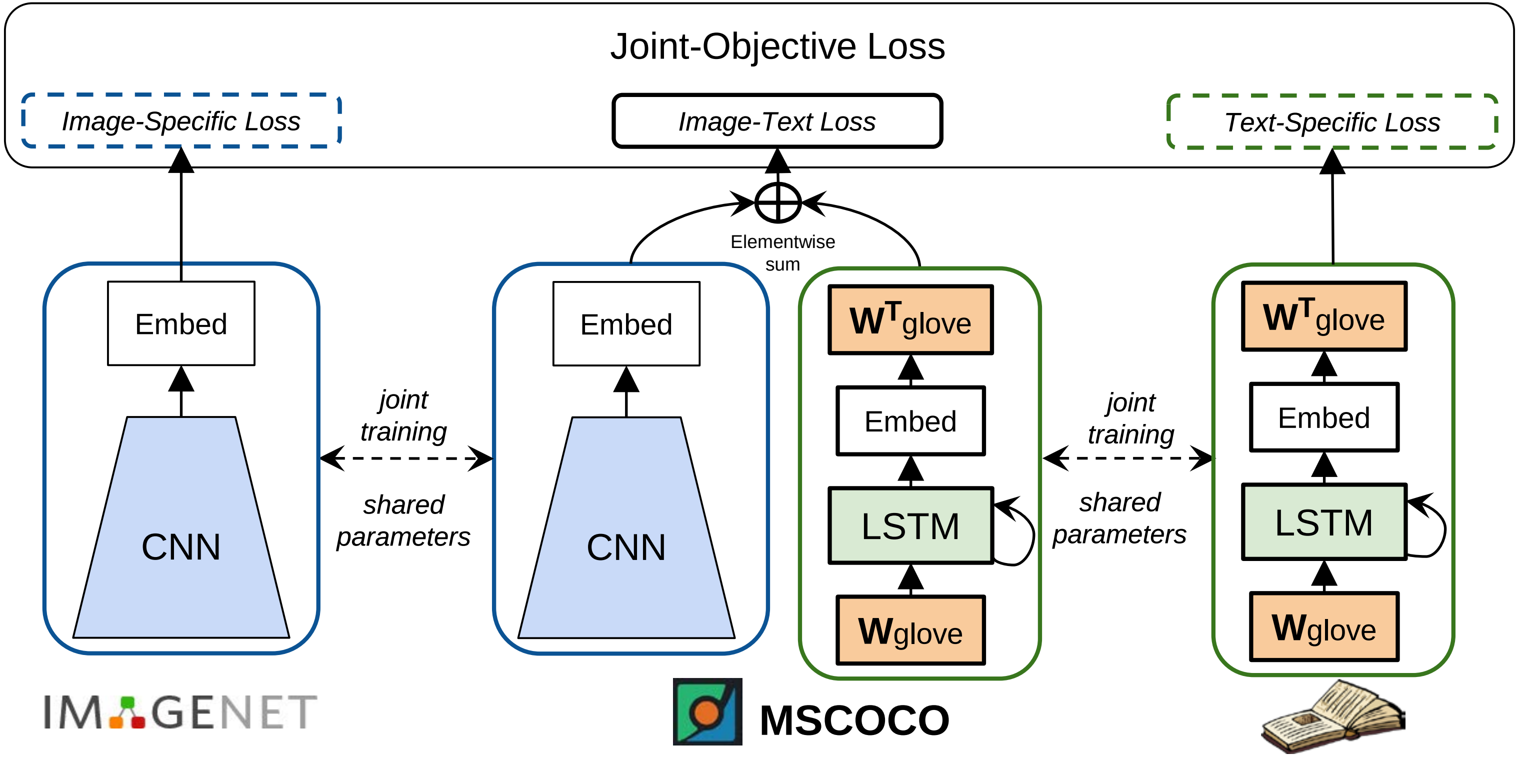
3. Overcome “forgetting” since pre-training alone is not sufficient.

[Catastrophic Forgetting in Neural Networks. Kirkpatrick et al. PNAS 2017]

Insight 3: Jointly train on multiple sources



Novel Object Captioner (NOC) Model



slide: Subhashini Venugopalan

Empirical Evaluation: COCO dataset In-Domain setting

MSCOCO Unpaired Image Data

Elephant, Galloping, Green, Grass

People, Playing, Ball, Field

Black, Train, Tracks

Eat, Pizza

Kitchen, Microwave

MSCOCO Paired Image-Sentence Data

"An elephant galloping in the green grass"

"Two people playing ball in a field"

"A black train stopped on the tracks"

"Someone is about to eat some pizza"

"A kitchen counter with a microwave on it"

MSCOCO Unpaired Text Data

"An elephant galloping in the green grass"

"Two people playing ball in a field"

"A black train stopped on the tracks"

"Someone is about to eat some pizza"

"A microwave is sitting on top of a kitchen counter "

Empirical Evaluation: COCO heldout dataset

MSCOCO Unpaired Image Data


Elephant, Galloping, Green, Grass


People, Playing, Ball, Field


Black, Train, Tracks


Pizza


Microwave

MSCOCO Paired Image-Sentence Data


"An elephant galloping in the green grass"


"Two people playing ball in a field"


"A black train stopped on the tracks"


"Someone is about to eat some pizza"


"A kitchen counter with a microwave on it"

Held-out

MSCOCO Unpaired Text Data

"An elephant galloping in the green grass"

"Two people playing ball in a field"

"A black train stopped on the tracks"

"A white plate topped with cheesy pizza and toppings."

"A white refrigerator, stove, oven dishwasher and microwave"

Empirical Evaluation: COCO

MSCOCO Unpaired Image Data

Two, elephants, Path, walking

Baseball, batting, boy, swinging

Black, Train, Tracks

Pizza

Microwave

MSCOCO Paired Image-Sentence Data

"An elephant galloping in the green grass"

"Two people playing ball in a field"

"A black train stopped on the tracks"

MSCOCO Unpaired Text Data

"A small elephant standing on top of a dirt field"

"A hitter swinging his bat to hit the ball"

"A black train stopped on the tracks"

"A white plate topped with cheesy pizza and toppings."

"A white refrigerator, stove, oven dishwasher and microwave"

- CNN is pre-trained on ImageNet

Empirical Evaluation: Metrics

F1 (Utility)**:** Ability to recognize and incorporate new words.
(Is the word/object mentioned in the caption?)

METEOR: Fluency and sentence quality.

Empirical Evaluation: Baselines

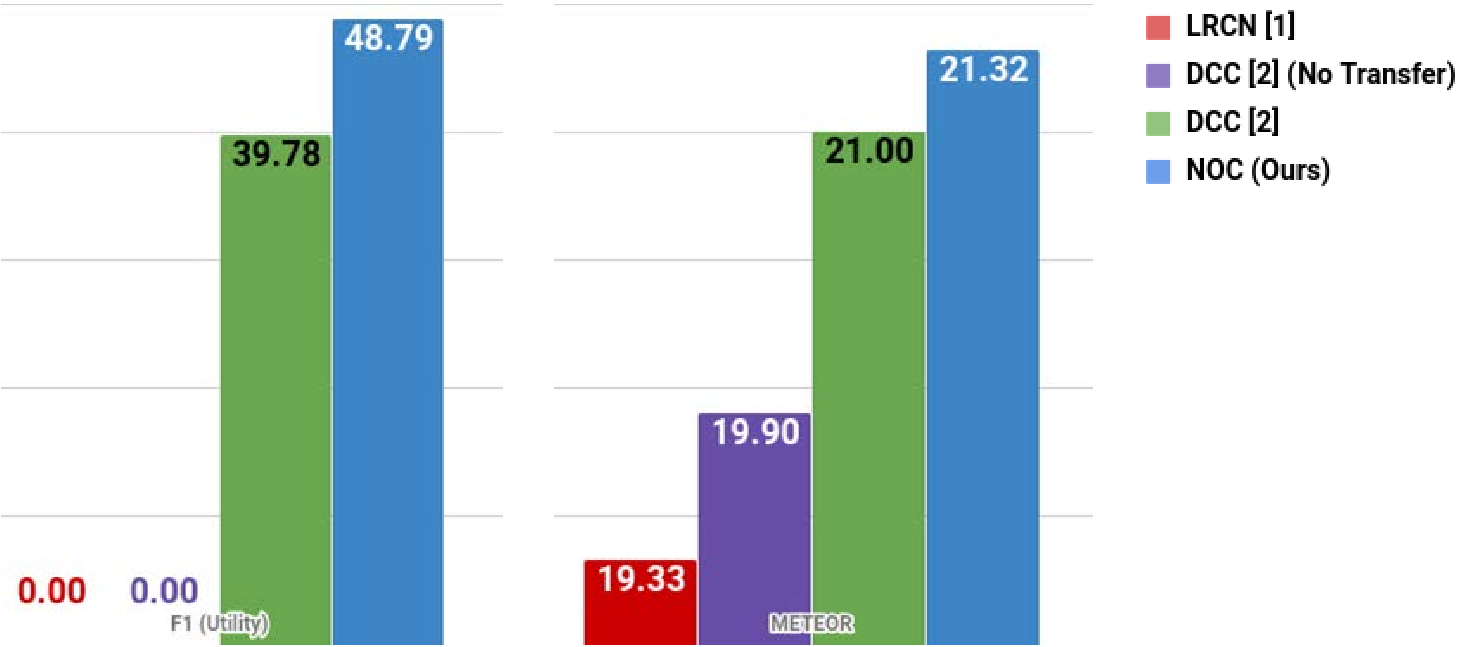
- LRCN [1]
- DCC [2] (No Transfer)
- DCC [2]
- NOC (Ours)

LRCN [1]: Does not caption novel objects.

DCC [2] : Copies parameters for the novel object from a similar object seen in training. (also not end-to-end)

[1] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell. CVPR'15
[2] L.A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, T. Darrell CVPR'16

Empirical Evaluation: Results



F1 (Utility)

METEOR (Fluency)

[1] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell. CVPR'15
[2] L.A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, T. Darrell CVPR'16

ImageNet: Human Evaluations

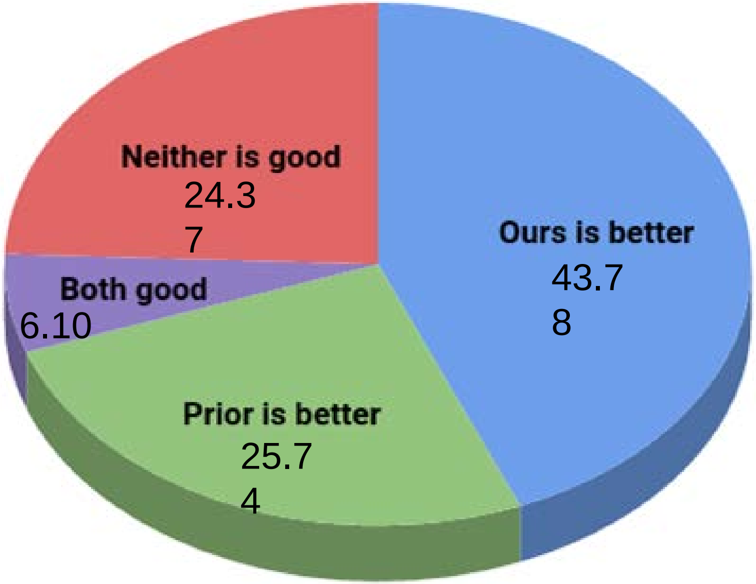
- **ImageNet:** 638 object classes not mentioned in COCO

NOC can describe 582 object classes
(60% more objects than prior work)

ImageNet: Human Evaluations

- **ImageNet:** 638 object classes not mentioned in COCO
- **Word Incorporation:** Which model incorporates the word (name of the object) in the sentence better?
- **Image Description:** Which sentence (model) describes the image better?

ImageNet: Human Evaluations



Word Incorporation

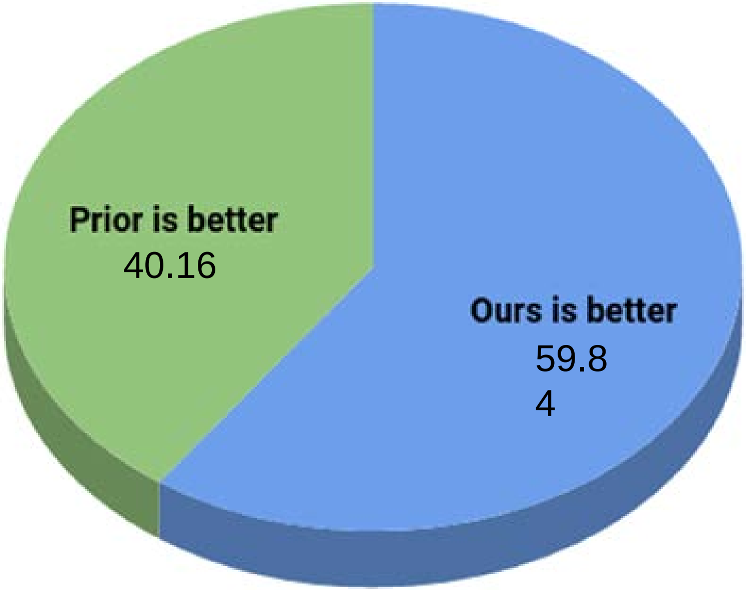


Image Description

Qualitative Evaluation: ImageNet

Instruments

A man holding a **banjo** in a park.

A large **chime** hanging on a metal pole

Vehicles

A **snowplow** truck driving down a snowy road.

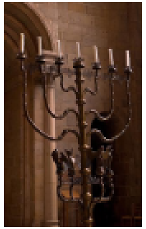
A group of people standing around a large white **warship**.

Land Animals

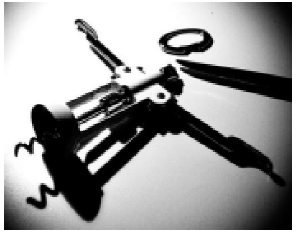
A **okapi** is in the grass with a **okapi**.

A small brown and white **jackal** is standing in a field.

Household



A large metal **candelabra** next to a wall.

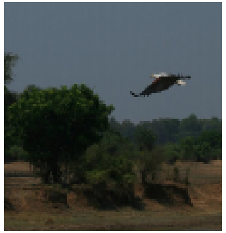


A black and white photo of a **corkscrew** and a **corkscrew**.

Qualitative Evaluation: ImageNet

Birds

A small **pheasant** is standing in a field.



A **osprey** flying over a large grassy area.

Outdoors

A large **glacier** with a mountain in the background.

A group of people are sitting in a **baobab**.

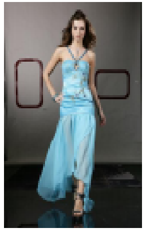
Water Animals

A **humpback** is flying over a large body of water.

A man is standing on a beach holding a **snapper**.

Misc

A table with a **cauldron** in the dark.



A woman is posing for a picture with a **chiffon** dress.

Qualitative Examples: Errors

Balaclava (n02776825) Error: Repetition
NOC: A **balaclava** black and white photo of a man in a **balaclava**.

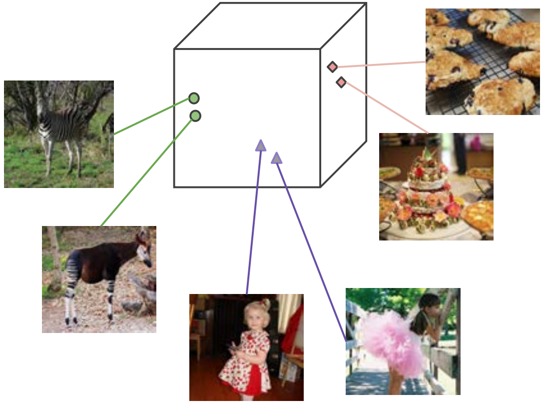
Sunglass (n04355933) Error: Grammar
NOC: A **sunglass** mirror reflection of a mirror in a mirror.

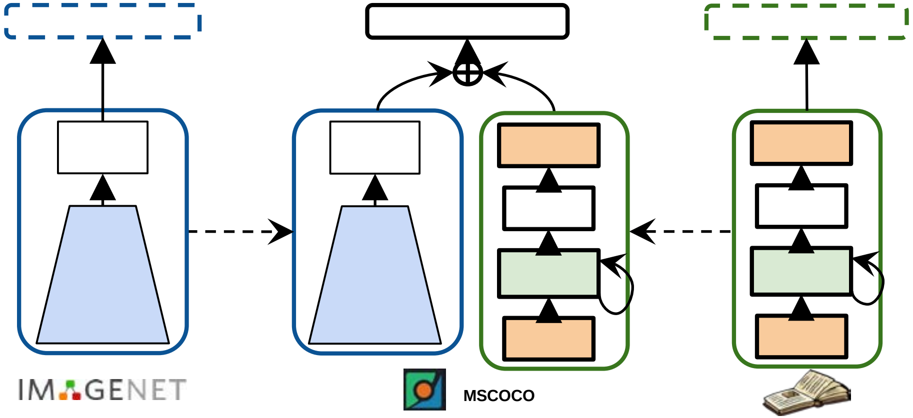
Gymnast (n10153594) Error: Gender, Hallucination
NOC: A man **gymnast** in a blue shirt doing a trick on a skateboard.

Cougar (n02125311) Error: Description
NOC: A **cougar** with a **cougar** in its mouth.

Novel Object Captioner - Take away

Semantic embeddings and joint training to caption 100s of objects.





A **okapi** standing in the middle of a field.

Vasili Ramanishka

Abir Das

Jianming Zhang

Thanks!

Subhashini Venugopalan

Lisa Anne Hendricks

Marcus Rohrbach

Raymond Mooney

Trevor Darrell