Natural-Language Video Description with Deep Recurrent Neural Networks

June 2017

Subhashini Venugopalan
University of Texas at Austin

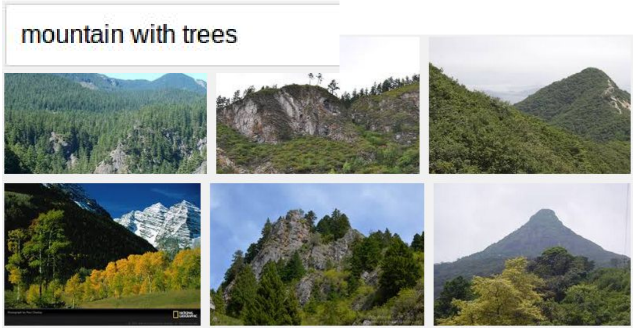
Problem Statement

Generate descriptions for events depicted in video clips

A monkey pulls a dog’s tail and is chased by the dog.

Applications

Image and video retrieval by content



Video description service

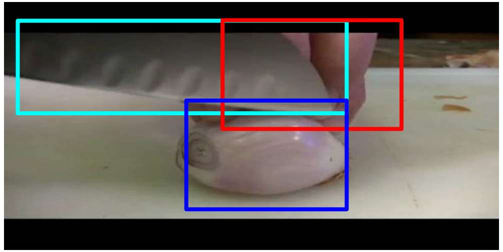
Human Robot Interaction

Video surveillance

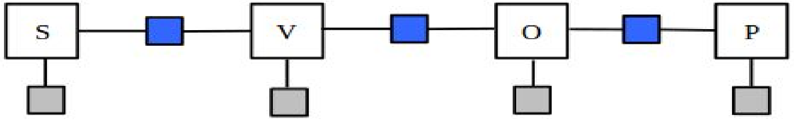
Outline

- Review (proposal)
    - Background
    - Encoder-Decoder approaches to video description
- External knowledge to improve video description
- External knowledge for novel object captioning
- Temporal segmentation and description for long videos
- Future Directions

Early Work in Video Description



Subjects		Verbs		Objects		Scenes	
person	0.95	slice	0.19	egg	0.31	kitchen	0.64
monkey	0.01	chop	0.11	onion	0.21	sky	0.17
animal	0.01	play	0.09	potato	0.20	house	0.07
.		.		.		.	
parrot	0	speak	0	piano	0	snow	0



A person is slicing an onion in the kitchen.

- Extract features
- Classify objects, actions, scenes

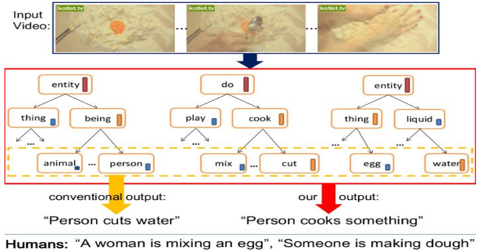
- Visual confidences over entities : Subject, Verb, Object, Scene

- Bias with statistics from language
- Factor Graph to estimates most likely entities (S, V, O, P)

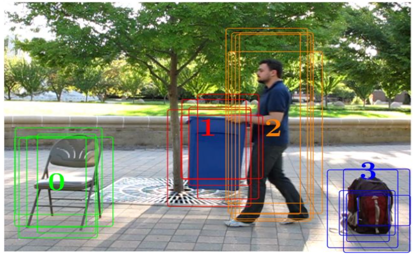
- Template based sentence generation.

J. Thomason*, S. Venugopalan*, S. Guadarrama, K. Saenko, R. Mooney COLING’14

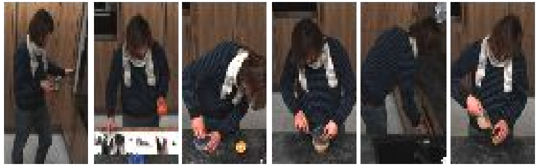
Early Work in Video Description



[Guadarrama, et al. ICCV'13]



[Yu and Siskind, ACL'13]



[Rohrbach et al. ICCV'13]

Limitations:

- Narrow Domains
- Small Grammars
- Template based sentences
- Several features and classifiers

Which objects/actions/scenes should we build classifiers for?

[Thomason et al. COLING'14]

Can we learn directly from video sentence pairs?

Without having to explicitly identify objects/actions/scenes to build classifiers.

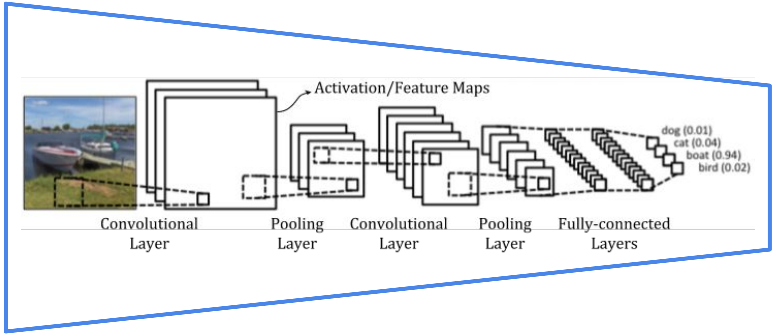
S. Venugopalan, H. Xu, M. Rohrbach, J. Donahue, R. Mooney, K. Saenko. NAACL'15

Outline

- Review (proposal)
  - Background
  - Encoder-Decoder approaches to video description
- External knowledge to improve video description
- External knowledge for novel object captioning
- Temporal segmentation and description for long videos
- Future Directions

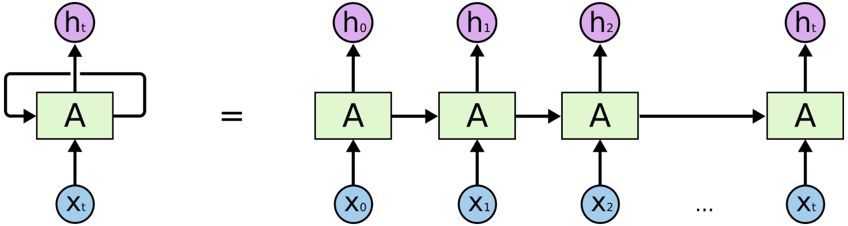
Deep Neural Networks

Convolutional Neural Networks



- Features and classifiers are jointly learned.
- Directly from raw pixels and labels.

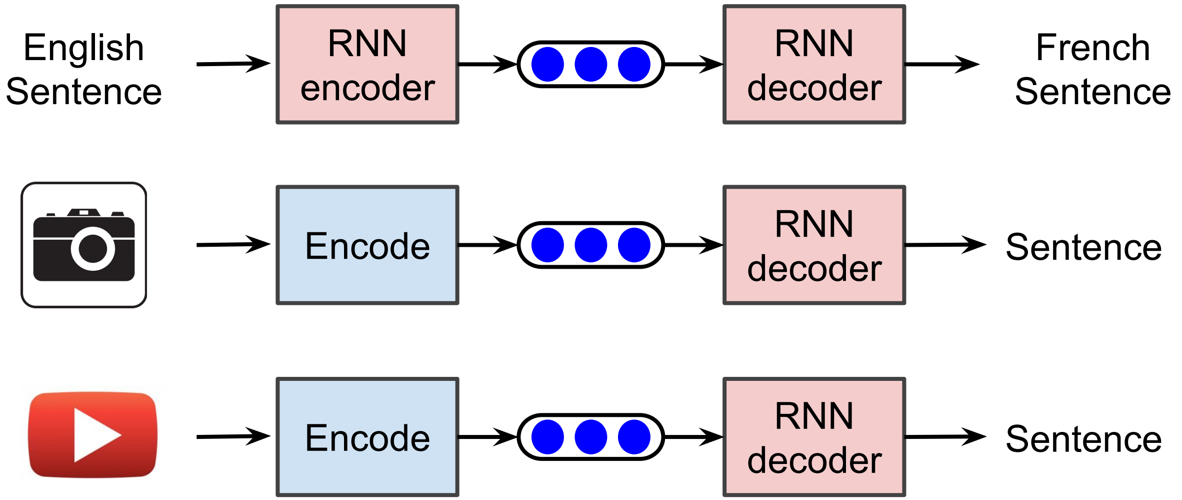
Recurrent Neural Networks



- RNNs can model sequences.
- Maps $(x_t, h_t) \rightarrow (y_t, h_{t+1})$

- Successful in translation, speech.
- We use LSTMs.

Recurrent Neural Networks (RNNs) can map a vector to a sequence.



[Sutskever et al. NIPS’14]

[Donahue et al. CVPR’15]

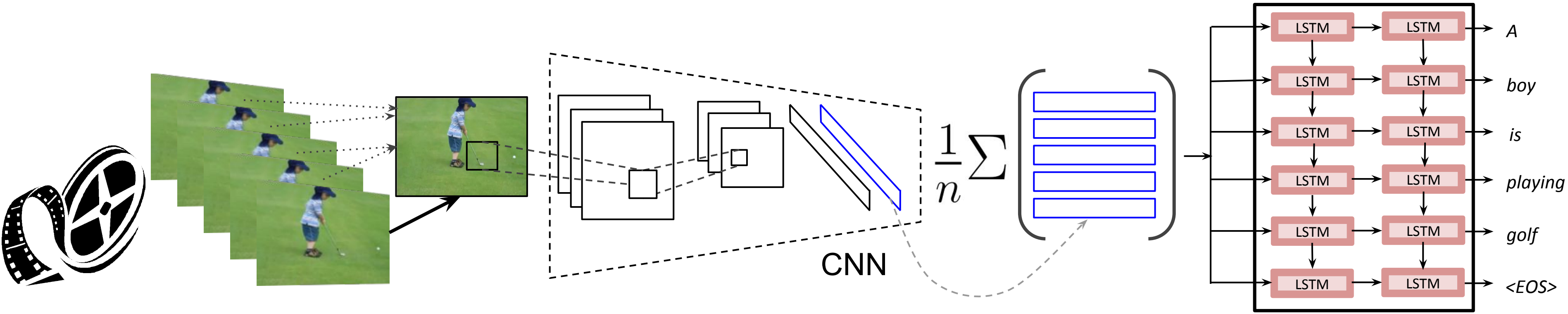
[Vinyals et al. CVPR’15]

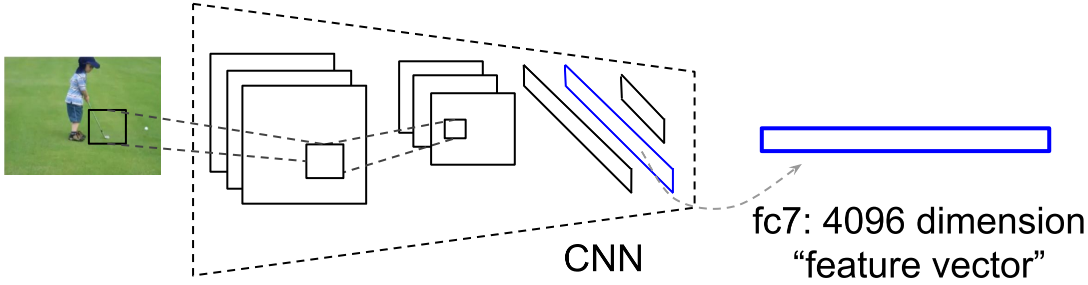
[Venugopalan et al. NAACL’15]

Key Insight:

Generate feature representation of the video and “decode” it to a sentence

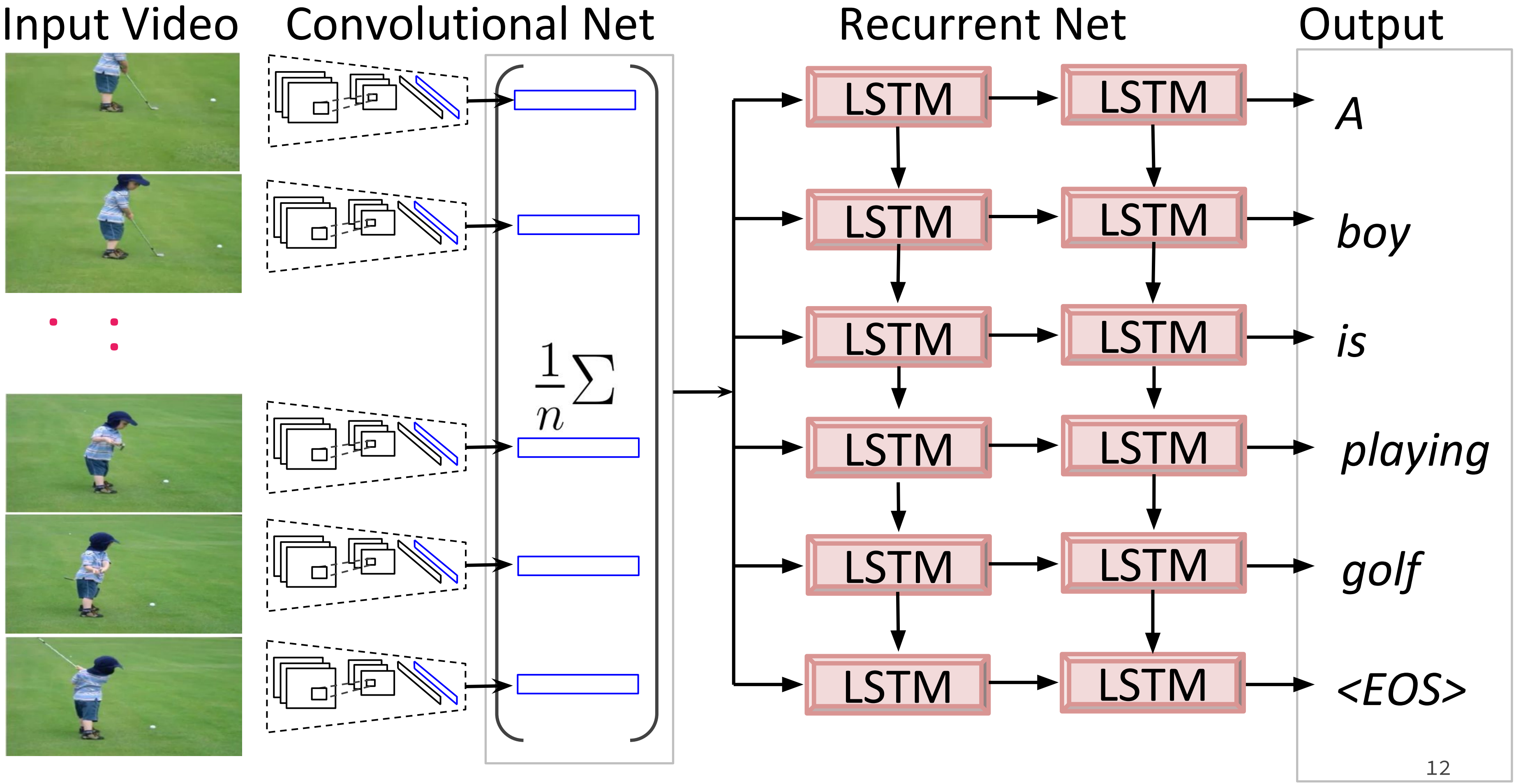
Inference: Feature extraction



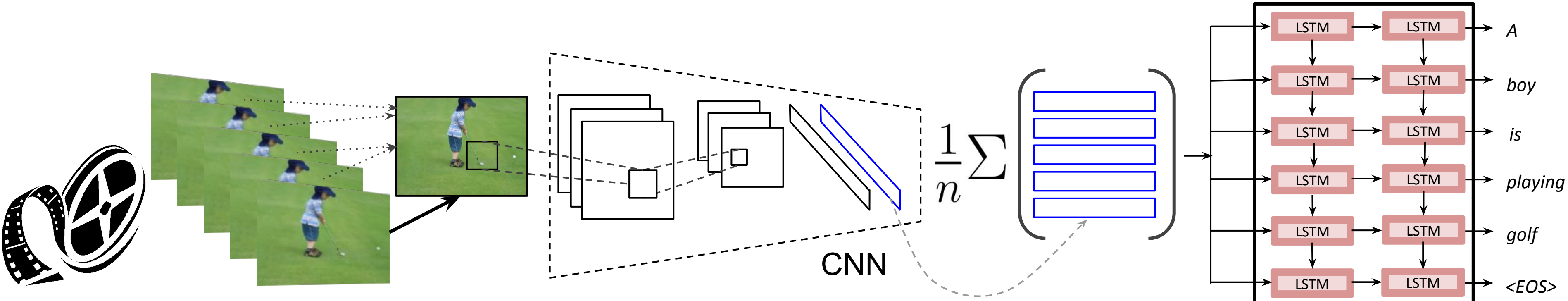


Forward propagate
Output: “fc7” features
(activations before classification layer)

Inference: Mean Pool & Generation

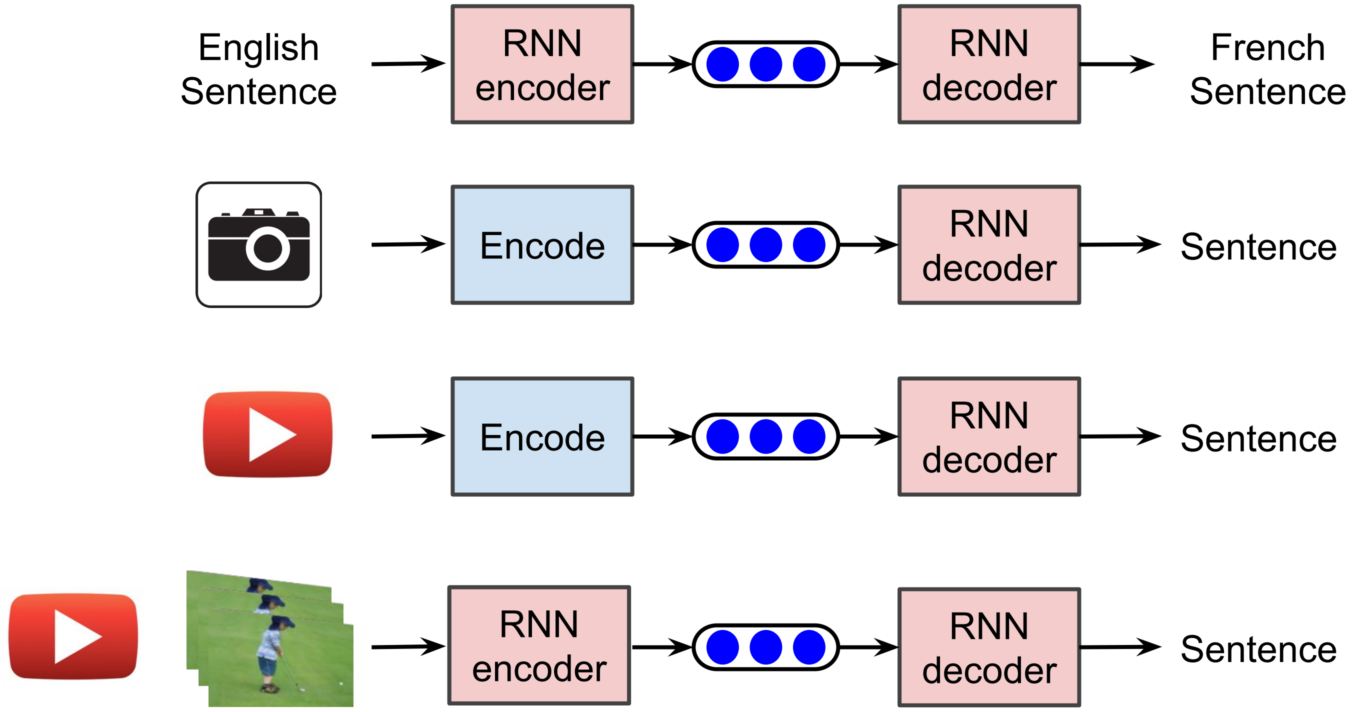


Translating Videos to Natural Language



Does not consider temporal sequence of frames.

Recurrent Neural Networks (RNNs) can map a vector to a sequence.



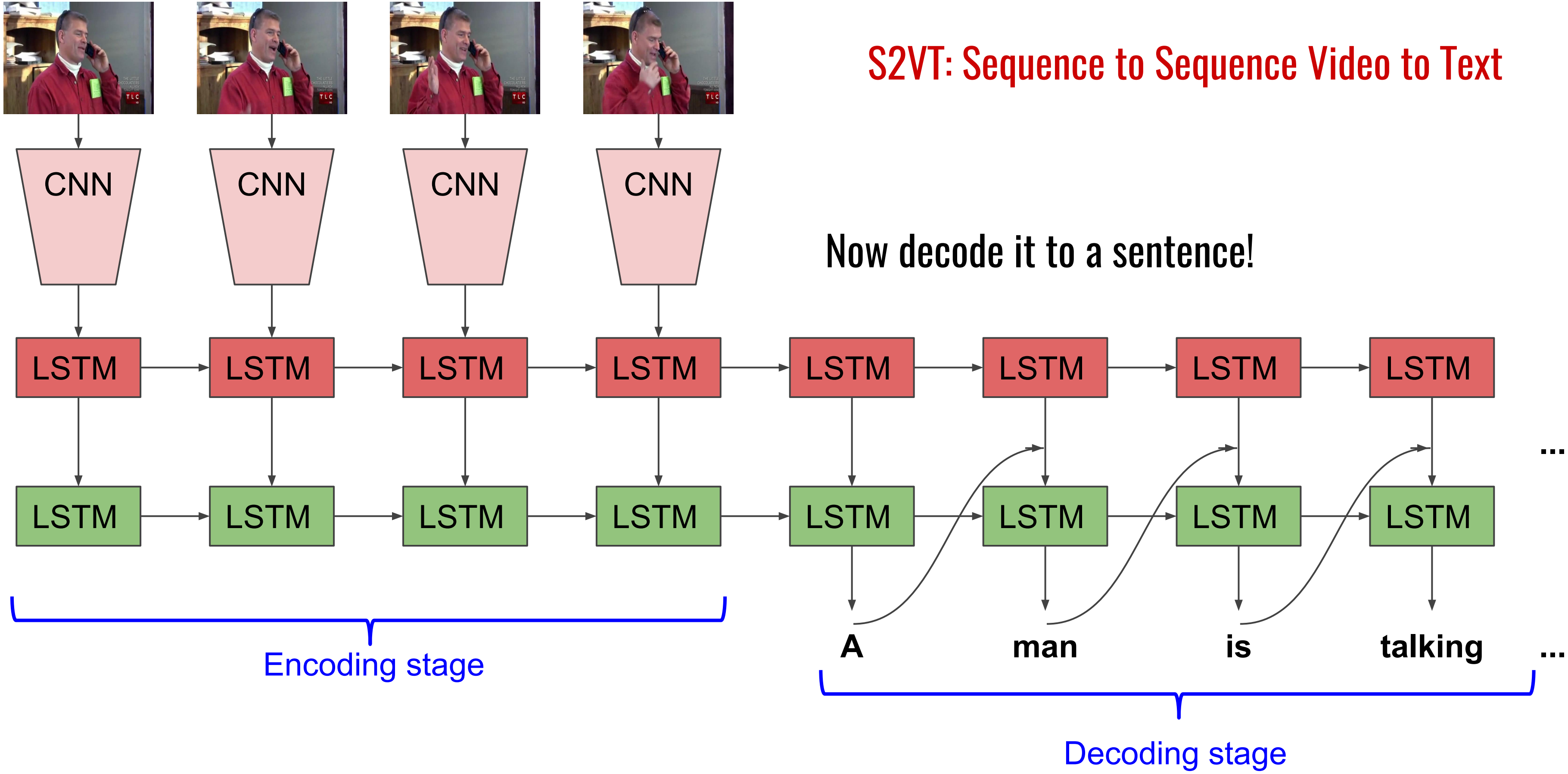
[Sutskever et al. NIPS'14]

[Donahue et al. CVPR'15]
[Vinyals et al. CVPR'15]

[Venugopalan et al. NAACL'15]

[Venugopalan et al. ICCV'15]

S2VT: Sequence to Sequence Video to Text

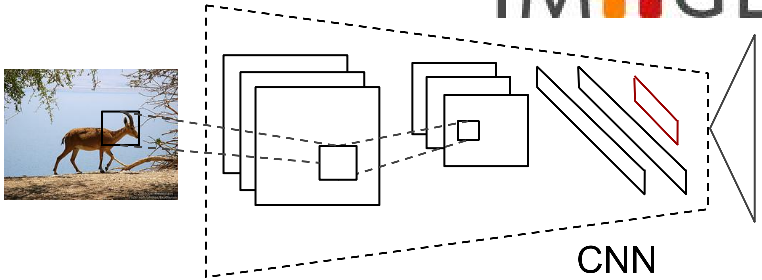


S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko. ICCV'15

Frames: RGB, Flow

IMAGENET

1. RGB frames.

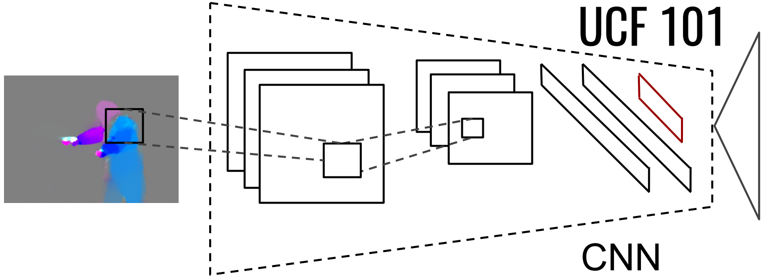


[Simonyan and Zisserman ICLR’15]

2. Use optical flow to extract flow images.

[T. Brox et al. ECCV ‘04]

3. Train CNN on Activity classes



[Donahue et al. CVPR’15]

Experiments: Dataset

Microsoft Research Video Description dataset [Chen & Dolan, ACL'11]

Link: http://www.cs.utexas.edu/users/ml/clamp/videoDescription/

- 1970 YouTube video snippets
  - 10-30s each
  - typically single activity
  - 1200 training, 100 validation, 670 test
- Annotations
  - Descriptions in multiple languages
  - ~40 English descriptions per video
  - descriptions and videos collected on AMT

Sample video and gold descriptions

- A man appears to be **plowing** a rice field with a plow being pulled by two **oxen**.
- A team of **water buffalo pull** a plow through a rice paddy.
- Domesticated **livestock** are helping a man **plow**.
- A man **leads** a team of **oxen** down a muddy path.
- Two **oxen walk** through some mud.
- A man is **tilling** his land with an **ox pulled** plow.
- **Bulls** are **pulling** an object.
- Two **oxen** are **plowing** a field.
- The farmer is **tilling** the soil.
- A man in **ploughing** the field.

- A man is **walking** on a **rope**.
- A man is **walking** across a **rope**.
- A man is **balancing** on a **rope**.
- A man is **balancing** on a **rope** at the beach.
- A man **walks** on a **tightrope** at the beach.
- A man is **balancing** on a **volleyball net**.
- A man is **walking** on a **rope** held by poles
- A man **balanced** on a **wire**.
- The man is **balancing** on the **wire**.
- A man is **walking** on a **rope**.
- A man is **standing** in the sea shore.

Movie Corpus - DVS

DVS - Separate audio track for the visually impaired

CC: Queen: “Which estate?”

DVS: Looking troubled, the Queen descends the stairs.

The Queen rushes into the courtyard. She then puts a head scarf on . . .

. . . and gets into the driver’s side of a nearby Land Rover.

The Land Rover pulls away.

Three bodyguards quickly jump into a nearby car and follow her.

Processed: Looking troubled, someone descends the stairs.

Someone rushes into the courtyard. She then puts a head scarf on ...

Evaluation: Movie Corpora

MPII-MD

- MPII, Germany
- DVS alignment: semi-automated and crowdsourced
- 94 movies
- 68,000 clips
- Avg. length: 3.9s per clip
- **~1 sentence per clip**
- 68,375 sentences

[Rohrbach et al. CVPR '15]

M-VAD

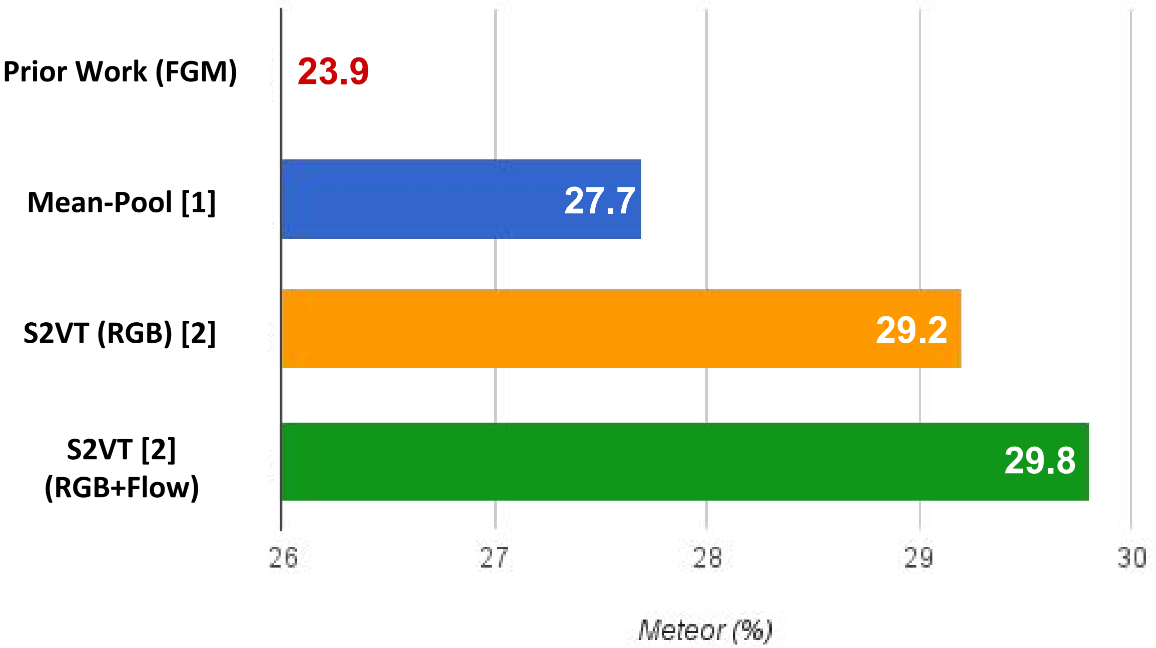
- Univ. of Montreal
- DVS alignment: semi-automated and crowdsourced
- 92 movies
- 46,009 clips
- Avg. length: 6.2s per clip
- **1-2 sentences per clip**
- 56,634 sentences

[Torabi et al. arXiv'15]

Evaluation Metrics

- Machine Translation Metric
  - METEOR - word similarity and phrasing
- Human evaluation
  - Relevance
  - Grammar

Results (Youtube)

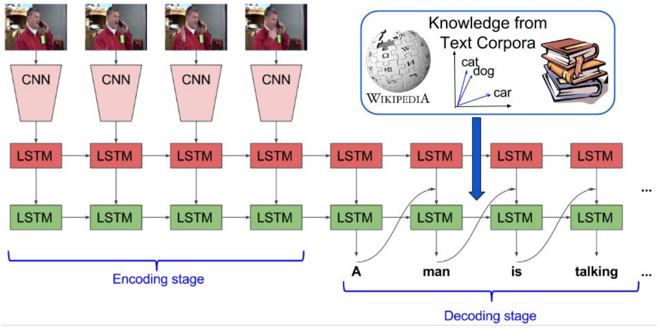


[1] S. Venugopalan, H. Xu, M. Rohrbach, J. Donahue, R. Mooney, K. Saenko. NAACL'15

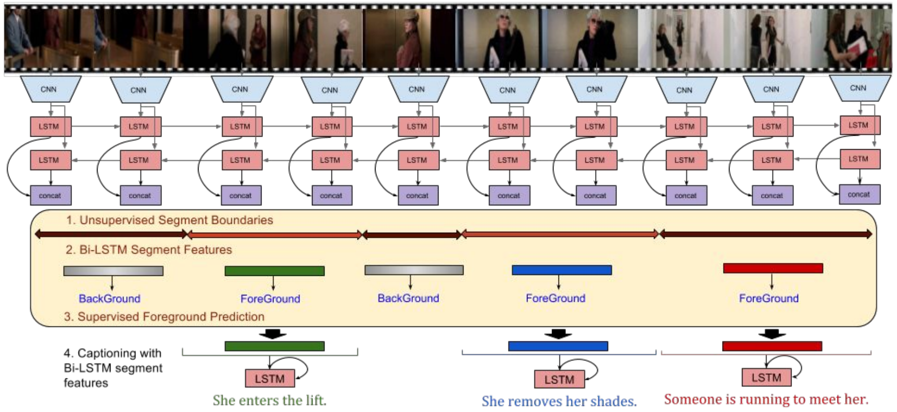
[2] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko. ICCV'15

Proposed Work

- Short Term - Incorporate linguistic knowledge to improve descriptions.



- Long Term - Descriptions for longer videos.



Outline

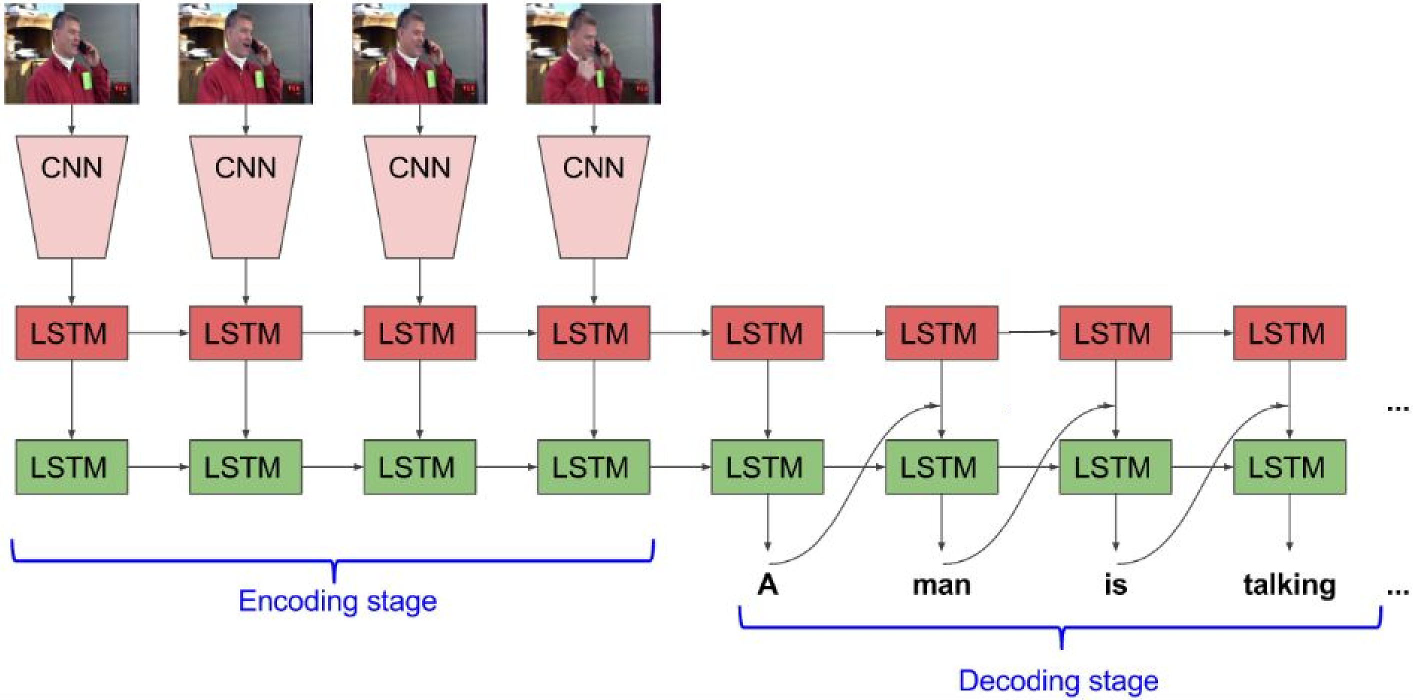
- Review (proposal)
  - Background
  - Encoder-Decoder approaches to video description
- External knowledge to improve video description
- External knowledge for novel object captioning
- Temporal segmentation and description for long videos
- Future Directions

Can external linguistic knowledge improve descriptive quality?

Unsupervised training on external text

S. Venugopalan, L.A. Hendricks, R. Mooney, K. Saenko. EMNLP'16

Integrating Statistical Linguistic Knowledge



S. Venugopalan, L.A. Hendricks, R. Mooney, K. Saenko. EMNLP'16

Unsupervised Training on External Text

Fusing LSTM language model trained on text

- Early fusion
- Late fusion
- Deep fusion

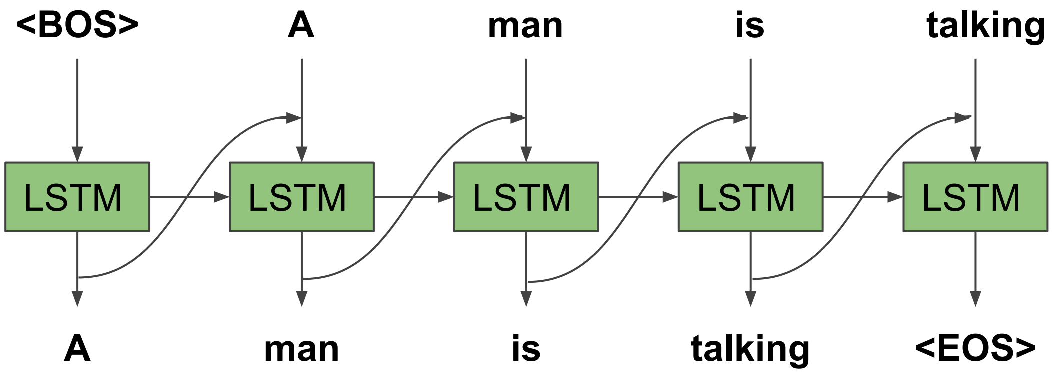
Distributional Embeddings

- Replace one-hot encoding with GloVe

LSTM Language Model

We learn a language model using LSTMs.

- Learns to predict the next word given previous words in the sequence.



- Data
  - Web Corpus: Wikipedia, UkWac, BNC, Gigaword
  - InDomain: MSCOCO image-caption sentences
  - Vocabulary: 72,700 (most frequent words)

Distributional Embedding

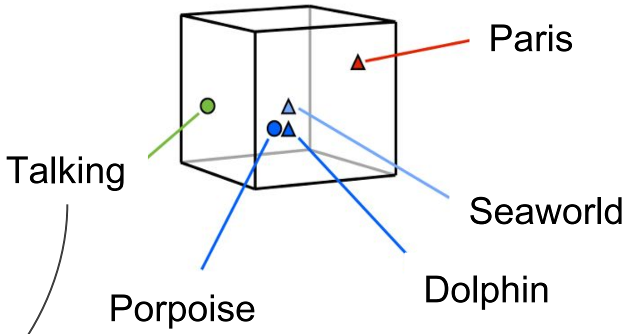
"You shall know a word by the company it keeps" (J. R. Firth, 1957)

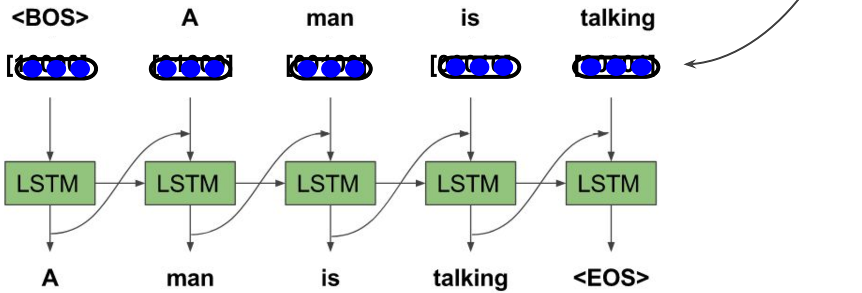
Dense vector representation of words.

- semantically similar words are closer.

We use GloVe [Pennington et al. EMNLP'14]

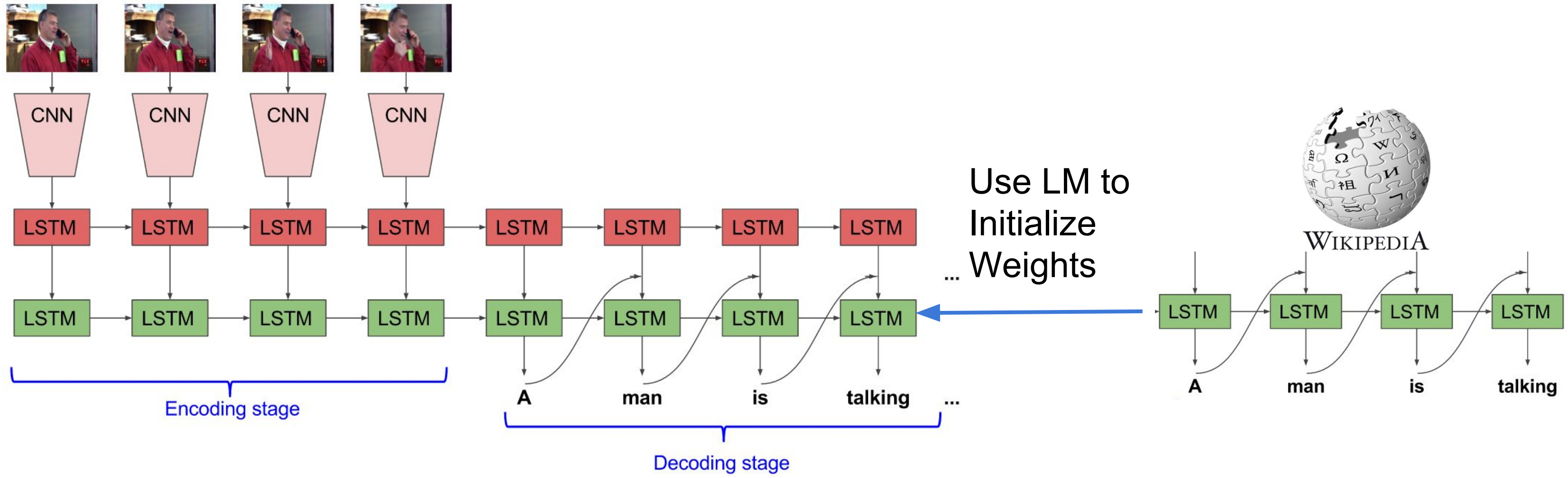
- Trained on Wikipedia and Gigaword. (6B tokens)
- Replace one-hot encoded input with GloVe.





Early Fusion

- Initialize weights of the caption model from the LSTM LM.

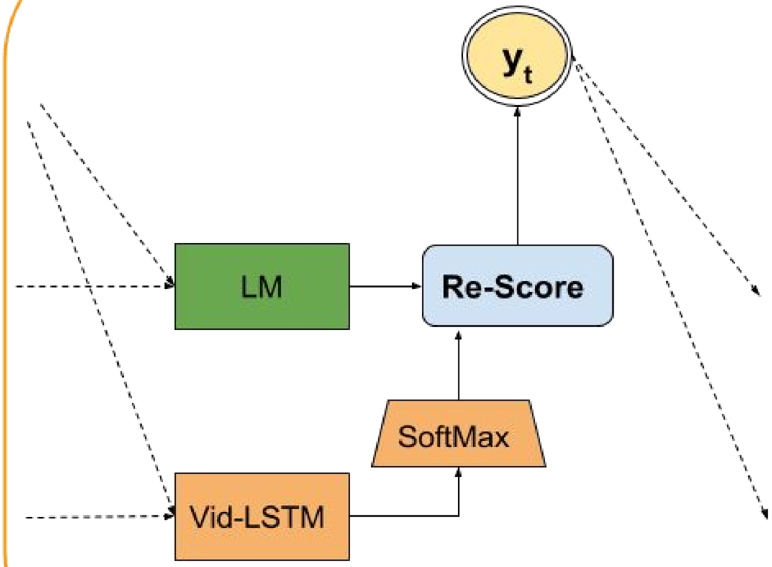


Late Fusion

Re-score video LSTM output based on language model.

$$p(y_t = y') = \alpha \cdot p_{VM}(y_t = y') + (1-\alpha) \cdot p_{LM}(y_t = y')$$

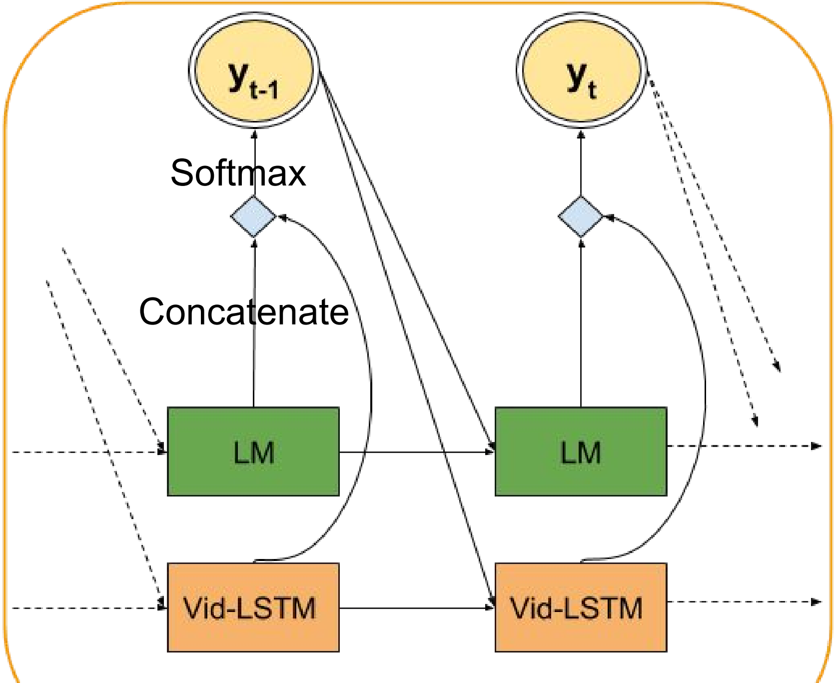
Set coefficient based on a validation set.



(a) Late Fusion

Deep Fusion

- Concatenate hidden states of LM LSTM and video caption LSTM.
- Fix LM, but train video caption model from scratch.
- Related MT work by [1]



(b) Deep Fusion

[1] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.C. Lin, F. Bougares, H. Schwenk,Y. Bengio. arXiv ‘15

Results (MSVD Dataset - Youtube clips)

Combining both techniques helps.

SOTA: HRNE (Pan. et al. CVPR’16) Hierarchical LSTM focuses on improving visual representation.
METEOR: 32.1 (no attn.), 33.1 (with attn.)

Human Evaluation

Relevance

Rate sentences based on how **accurately** they describe the event depicted in the video.

Least relevant Most Relevant

1 2 3 4 5

Grammar

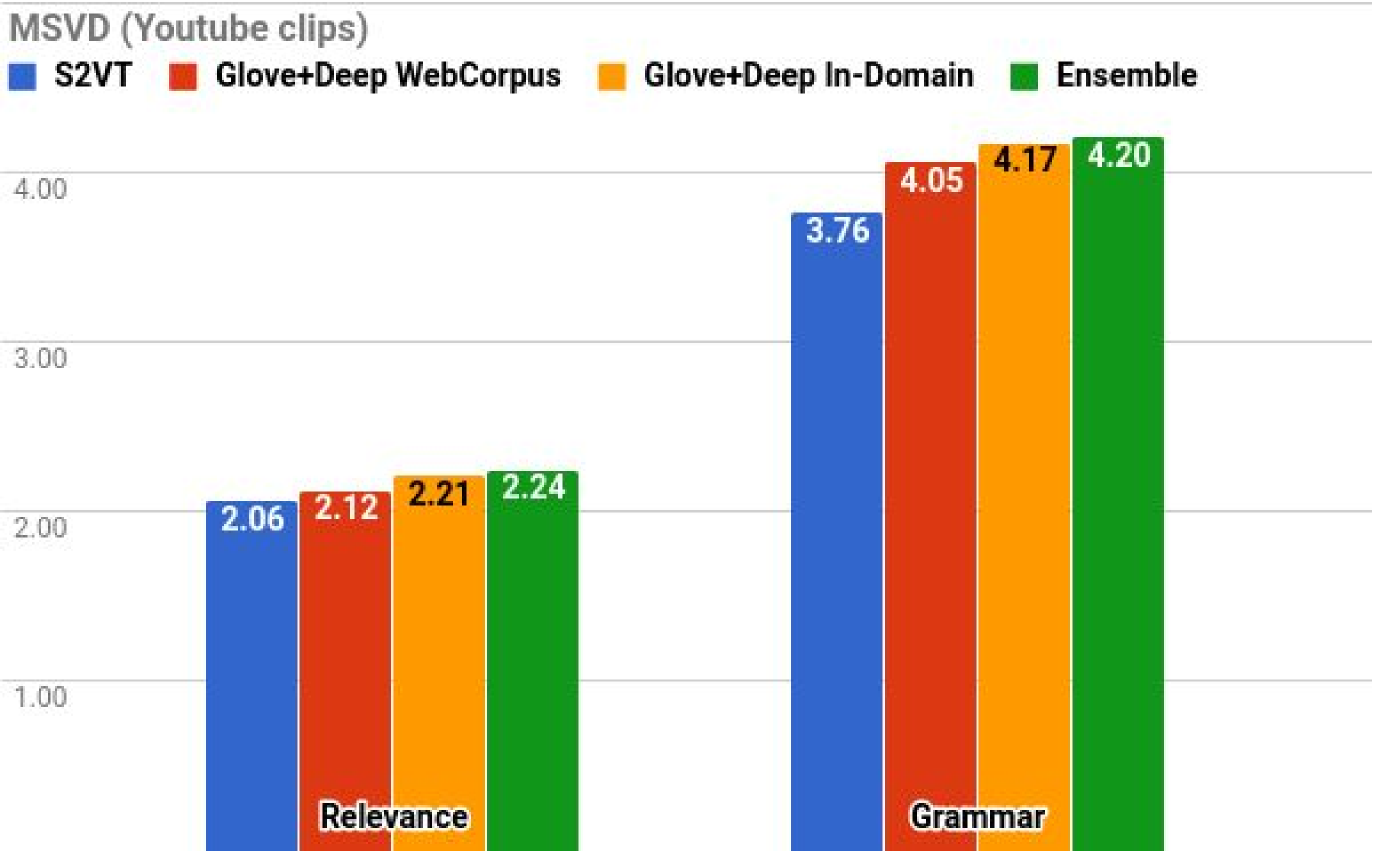
Rate the **grammatical correctness** of the following sentences.

Incorrect Grammatically correct

1 2 3 4 5

Sentences from the different models can have the same rating.

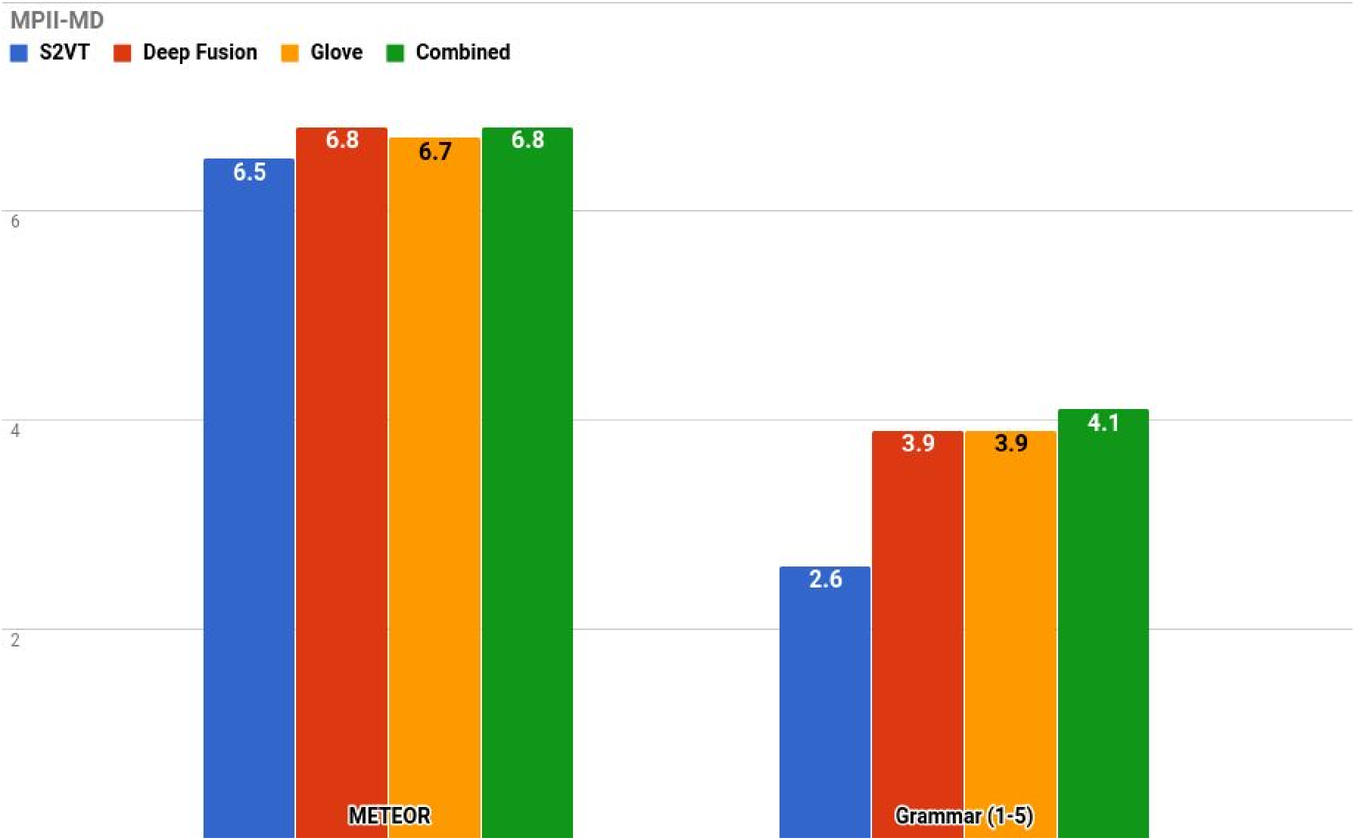
Results (Youtube) - Human Evaluation



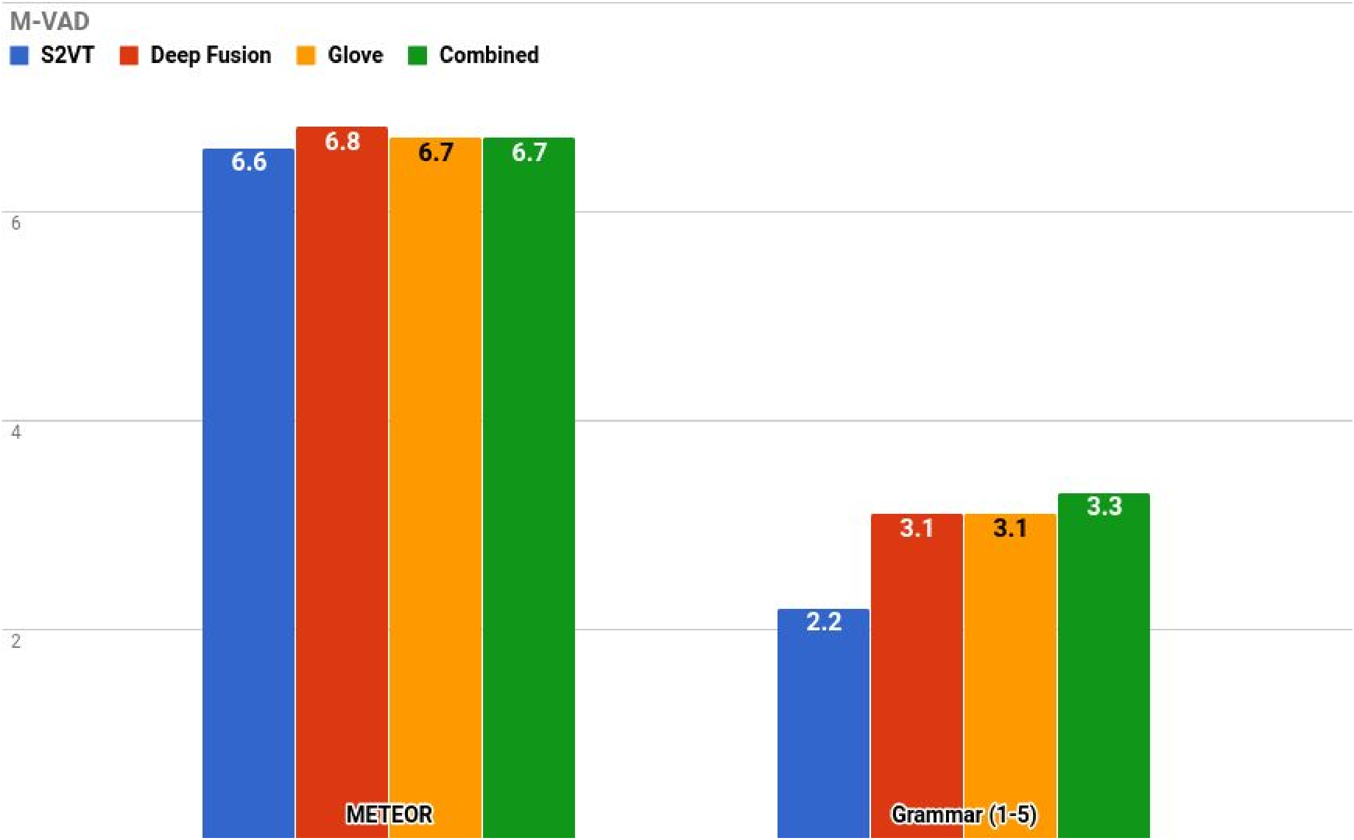
Examples

http://vsubhashini.github.io/language_fusion.html

Results - Movie Corpus (MPII-MD)



Results - Movie Corpus (M-VAD)



External knowledge can particularly help in captioning novel objects.

When there’s no paired training data.

S. Venugopalan, L.A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, K. Saenko. CVPR’17

Outline

- Review (proposal)
  - Background
  - Encoder-Decoder approaches to video description
- External knowledge to improve video description
- External knowledge for novel object captioning
- Temporal segmentation and description for long videos
- Future Directions

A brown bear walking across a lush green field.

A large brown bear walking through a forest.

A brown bear walks in the grass in front of trees.

A brown bear sitting on top of a green field.

A brown bear walking on a grassy field next to trees.

A large brown bear walking across a lush green field.

Image Credit: L.A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, T. Darrell, K. Saenko. CVPR'16

Novel Object Captioner

We present Novel Object Captioner which can compose descriptions about novel objects in context.

NOC (ours): Jointly train on **multiple sources** with **auxiliary objectives**.

IMAGENET + MSCOCO +

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

Visual Classifiers. IMAGENET

→ **okapi**

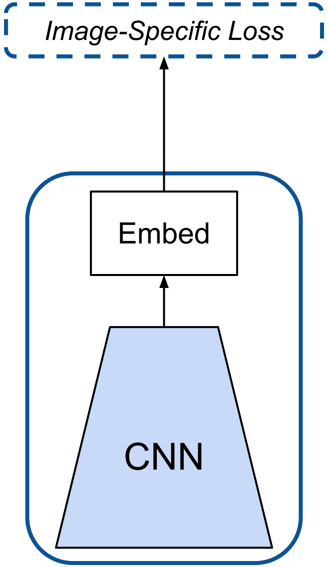
Existing captioners. IMAGENET init + train MSCOCO

→ A horse standing in the dirt.

S. Venugopalan, L.A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, K. Saenko. CVPR'17

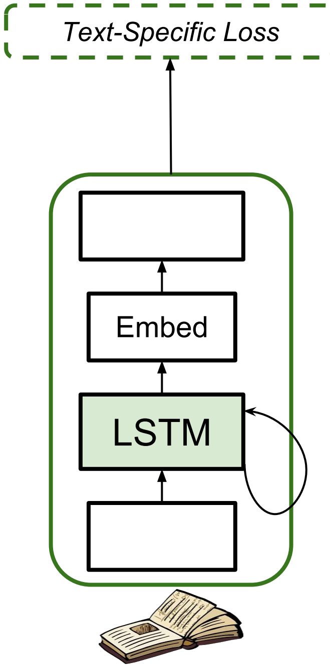
Key Insights

1. Train effectively on external sources



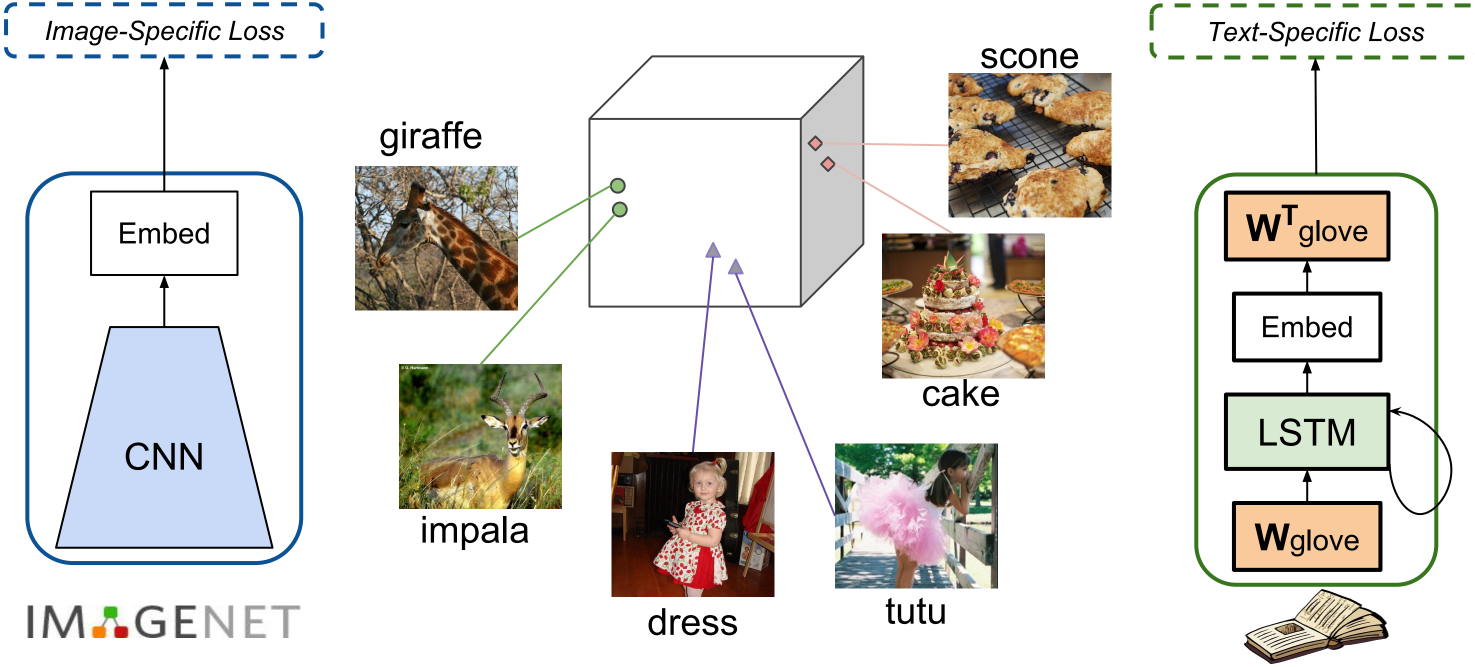
Visual features from unpaired image data

Language model from unannotated text data



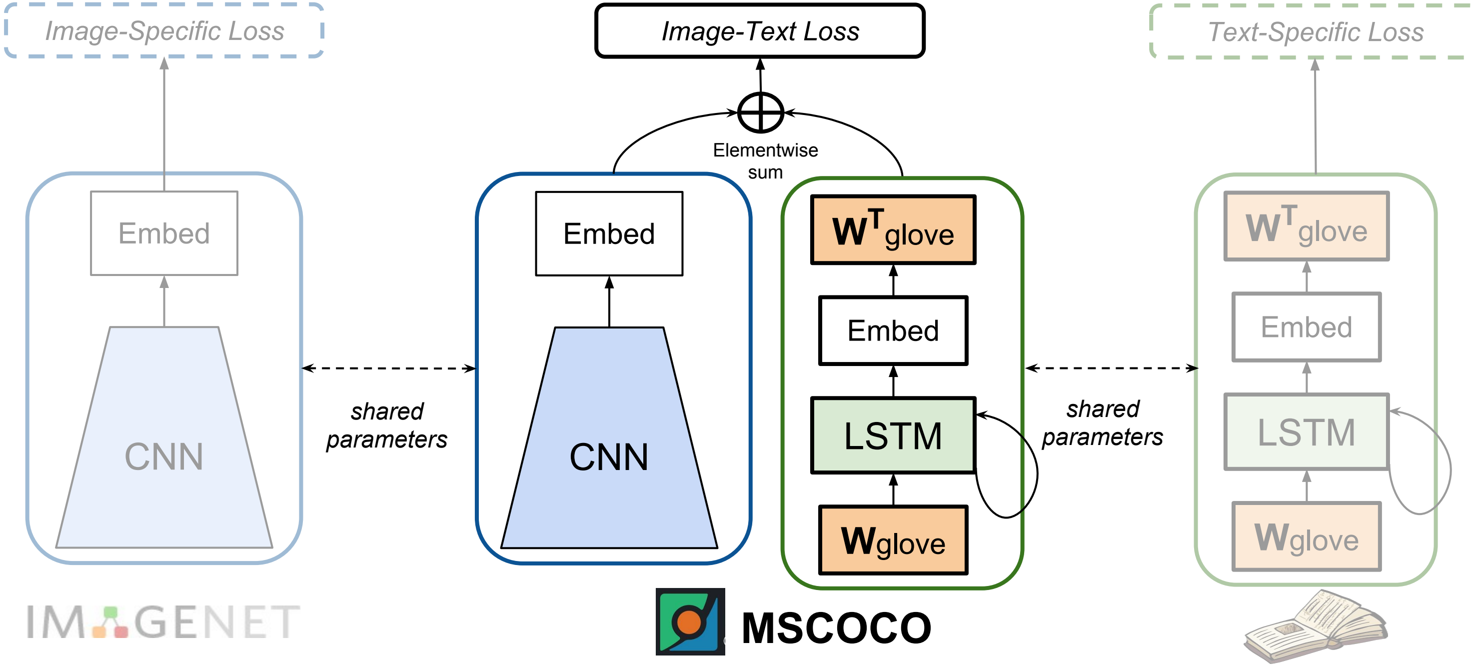
Key Insights

2. Capture semantic similarity of words



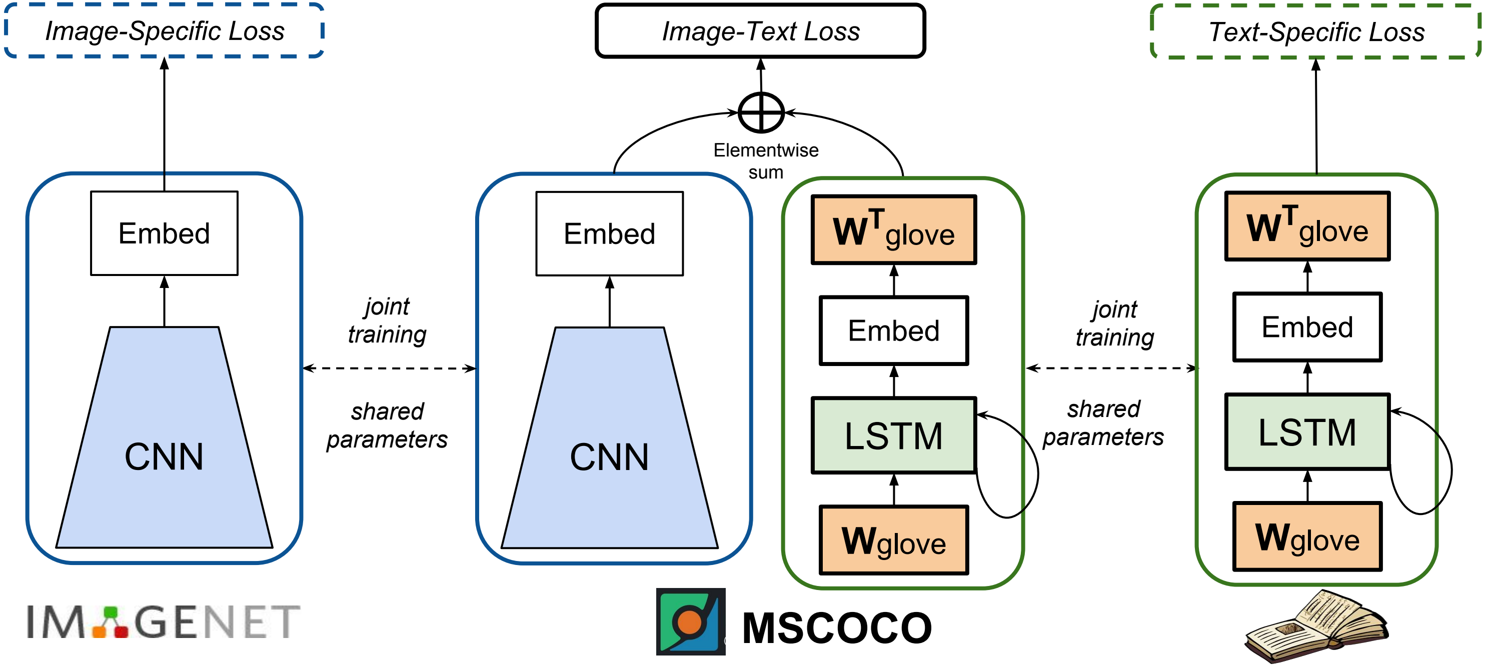
Key Insights

Combine to form a caption model

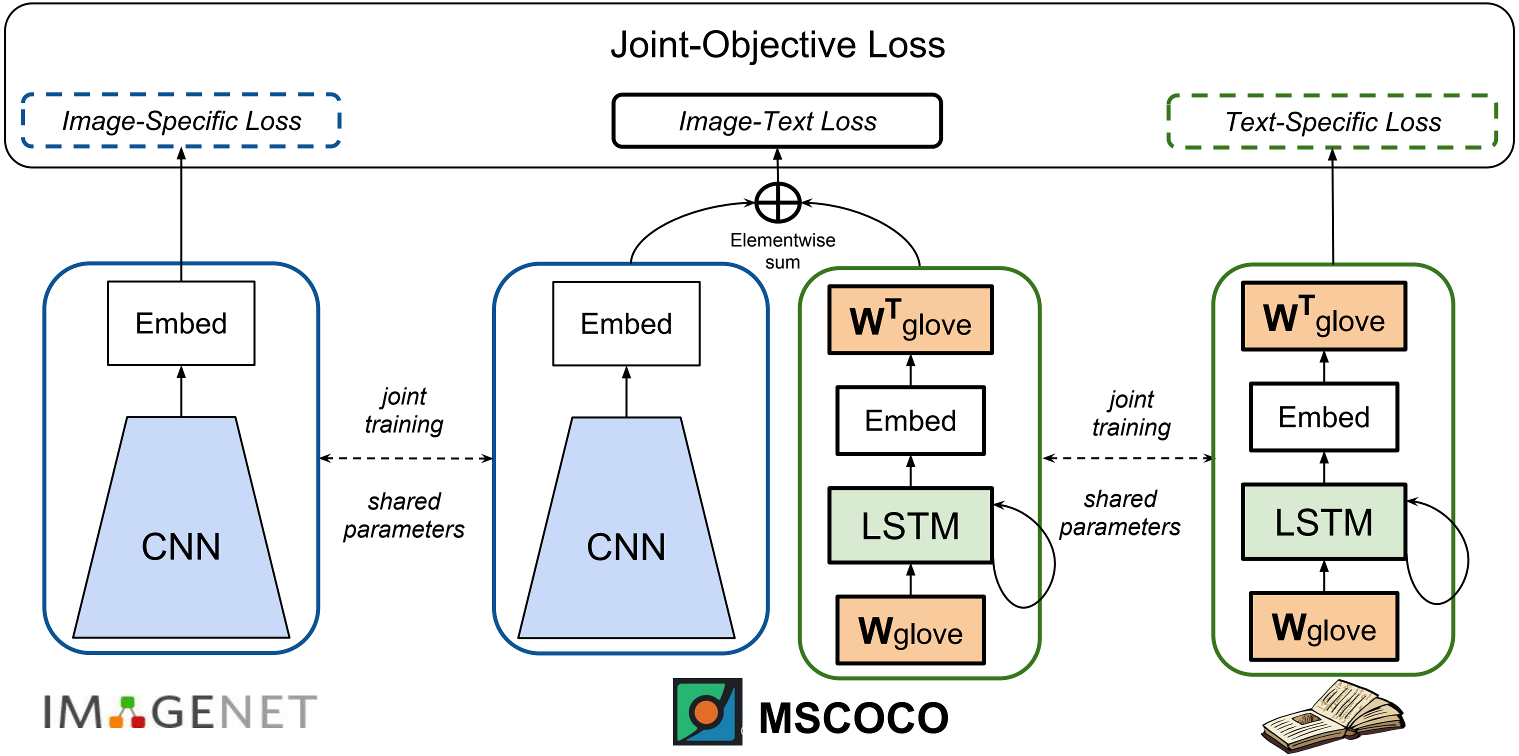


Key Insights

3. Jointly train on multiple sources

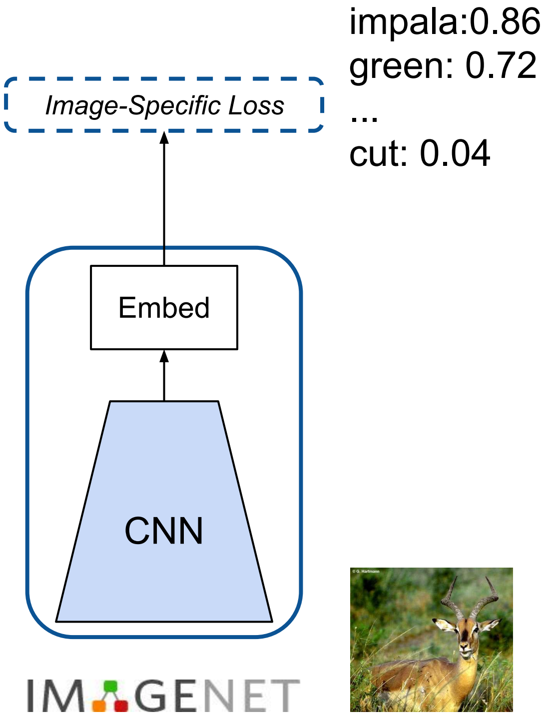


NOC Model



Joint-Objective Loss
Image-Specific Loss
Image-Text Loss
Text-Specific Loss
Elementwise sum
Embed
CNN
joint training
shared parameters
Embed
CNN
W^T glove
Embed
LSTM
W glove
joint training
shared parameters
W^T glove
Embed
LSTM
W glove
IMAGENET
MSCOCO

Visual Network



Network: VGG-16 with multi-label loss [sigmoid cross-entropy (logistic) loss]

Training Data: Unpaired image data

Features: Vector with activations corresponding to scores for *words in the vocabulary*.

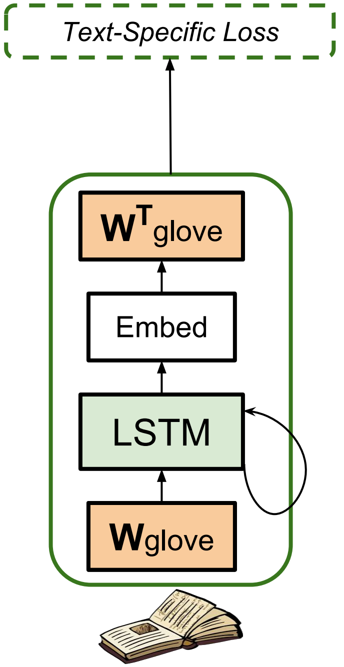
Language Model

Network: Pre-trained GloVe embeddings + LSTM layer. Predict a word w_{t+1} given previous words $w_{0..t}$ $p(w_{t+1} \mid w_{0..t})$

$(\mathbf{W}_{glove})^T$: Shared weights with input embedding.

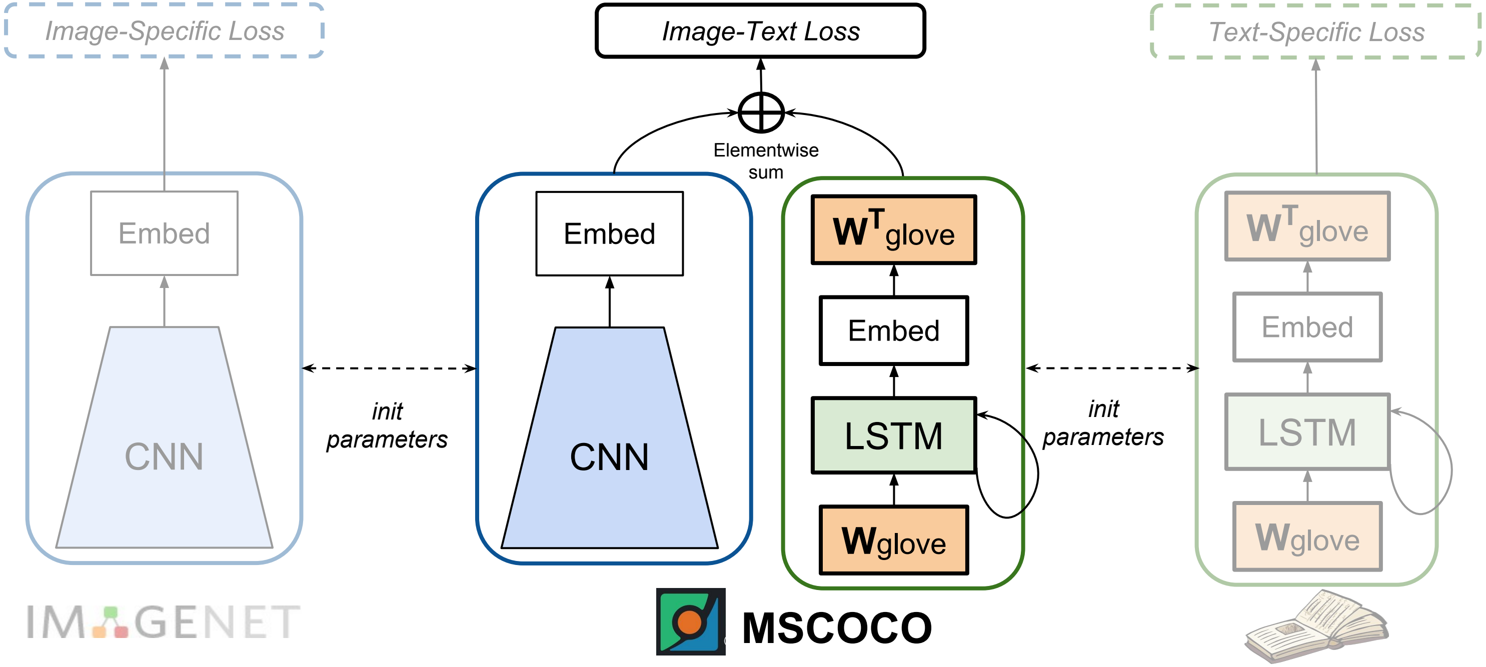
Training Data: Unannotated text data (BNC, ukWac, Wiki, Gigaword)

Features: Vector with activations corresponding to scores for *words in the vocabulary*.



Caption Model

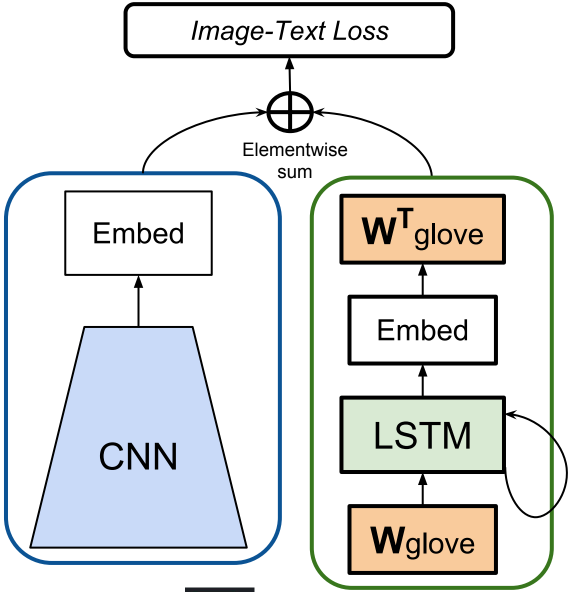
Network: Combine output of the visual and text networks. (softmax + cross-entropy loss)



Caption Model

Training Data:
COCO images with multiple labels

bear, brown, field, grassy, trees, walking

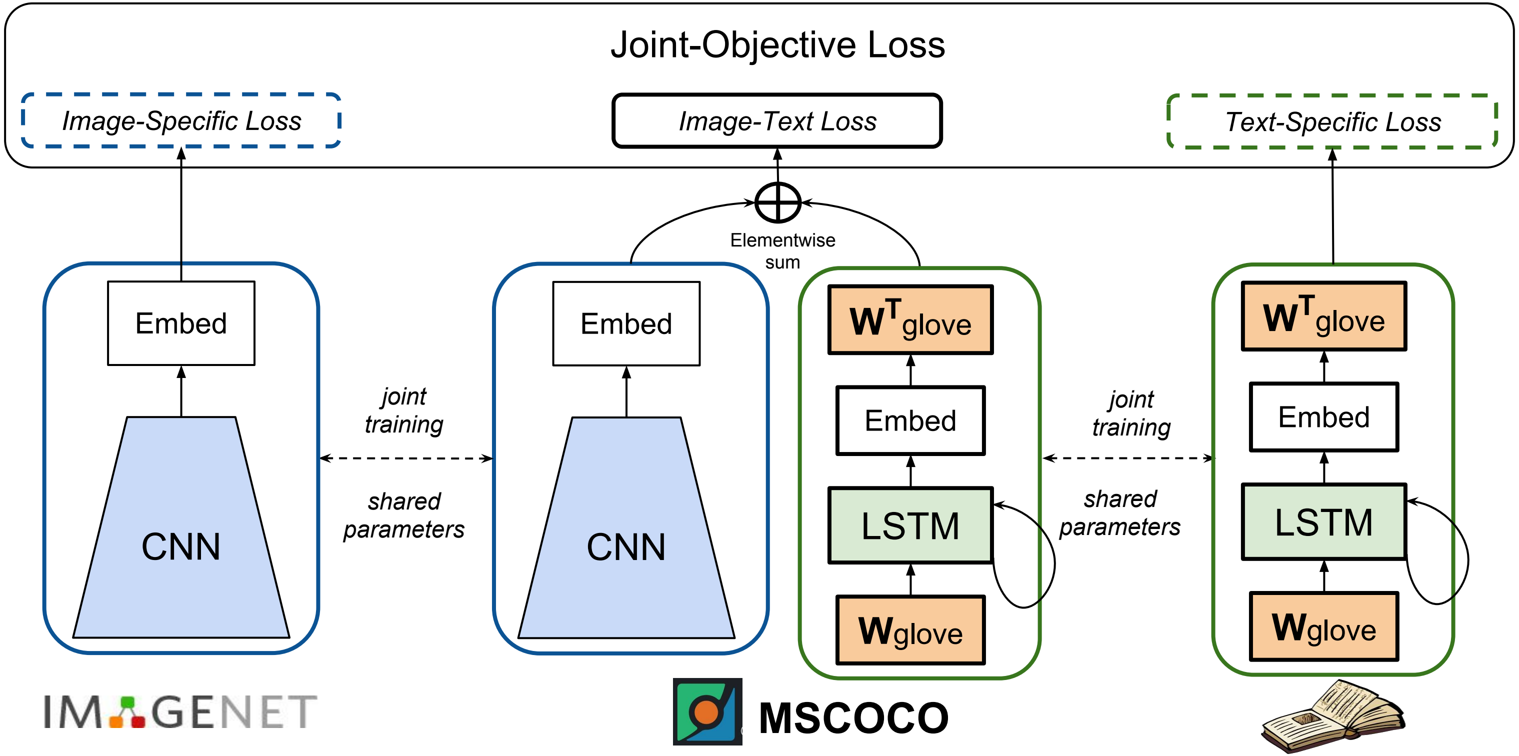


MSCOCO

Training Data:
Captions from MSCOCO

A brown bear walking on a grassy field next to trees

NOC Model: Train simultaneously



Evaluation

- Empirical: COCO held-out objects
  - **In-domain [Use images from COCO]**
  - Out-of-domain [Use imagenet images for same concepts]
- Ablations
  - Embedding & joint training contribution
- Human Evaluations: ImageNet
- Qualitative: ImageNet
  - **Objects not in COCO**
  - Rare objects in COCO

Empirical Evaluation: COCO dataset

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

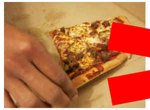

Microwave

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 cut a 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”

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

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

Held-out dataset

Empirical Evaluation: COCO In-Domain setting

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

Results: COCO In-Domain

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

METEOR: Fluency and sentence quality.

F1 (Utility)

METEOR

Results: COCO In-Domain

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

LRCN [1]: Does not caption novel objects.

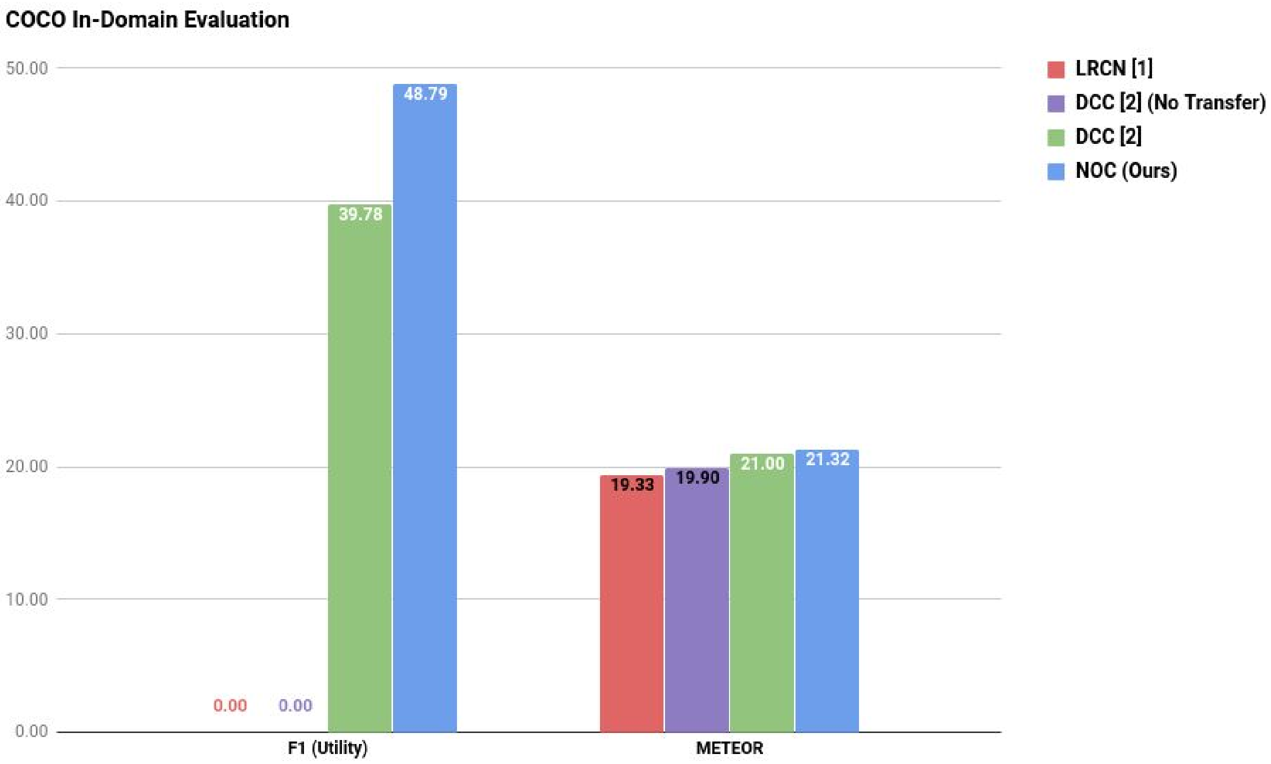
DCC [2] : Copies parameters for the novel object from a similar object seen in training.

F1 (Utility)

METEOR

[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

Results: COCO In-Domain



[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

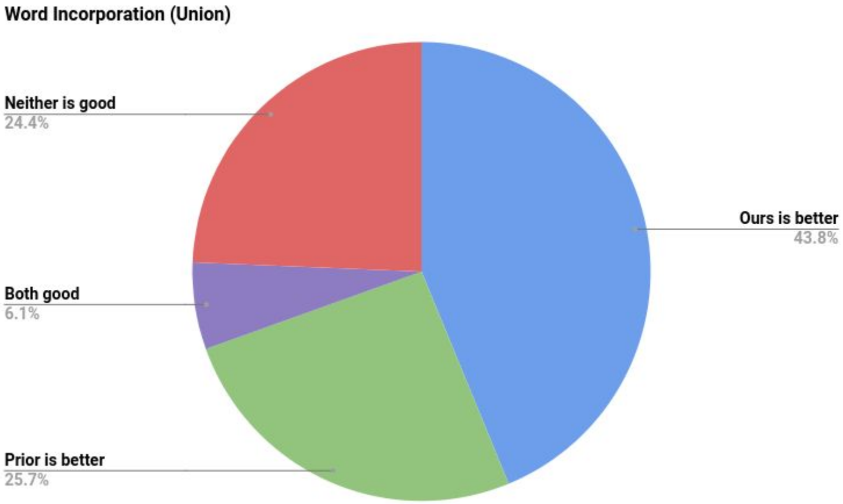
ImageNet: Human Evaluations

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

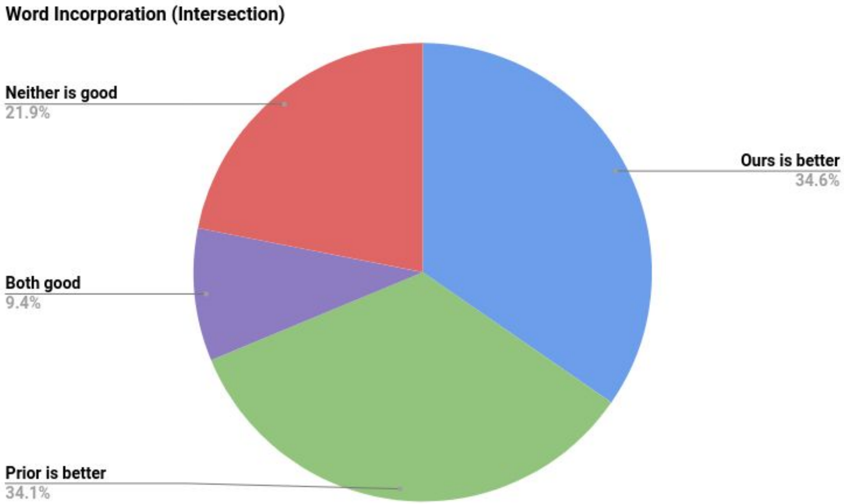
- **Union:** Objects that either model can describe.

- **Intersection:** Only the subset of objects that both models can describe. (~60%, ~380 categories)

ImageNet: Human Evaluations - Word Incorporation



Union

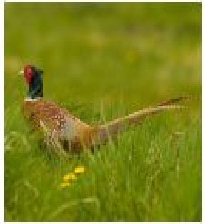


Intersection

Qualitative Evaluation: ImageNet

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.

Outline

- Review (proposal)
  - Background
  - Encoder-Decoder approaches to video description
- External knowledge to improve video description
- External knowledge for novel object captioning
- Temporal segmentation and description for long videos
- Future Directions

Localization and Description

Existing Video Captioning Methods

A woman is running down a corridor.

DVS Applications

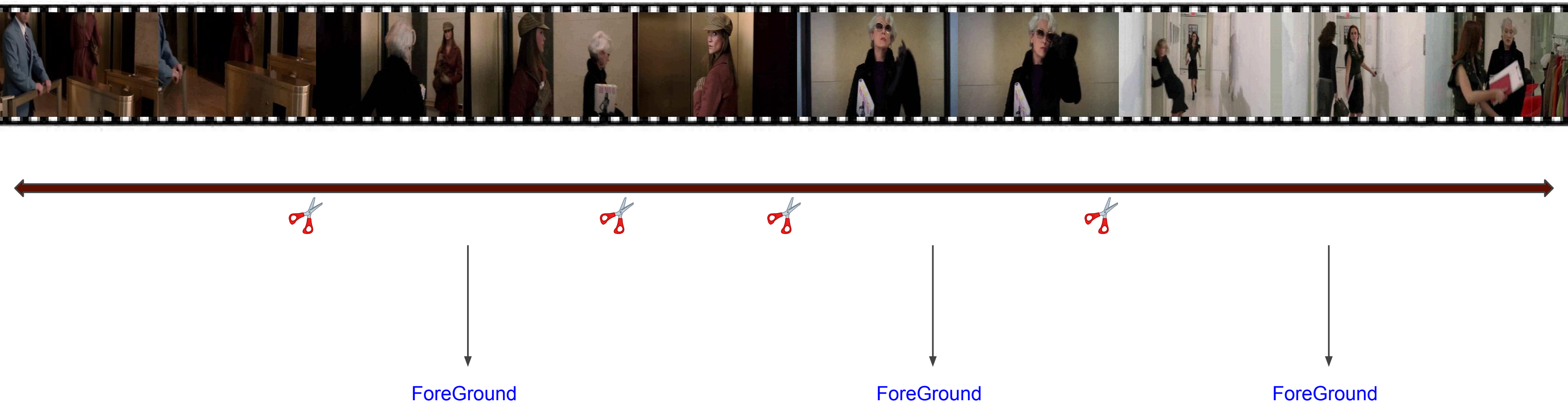
Someone strides through the foyer and approaches a lift.

The staff member waits for another lift.

She pulls off her designer shades.

Someone's running to meet her.

Overview



She enters the lift. She removes her shades. Someone is running to meet her.

S. Venugopalan, V. Ramanishka, M. Rohrbach, R. Mooney, T. Darrell, K. Saenko.

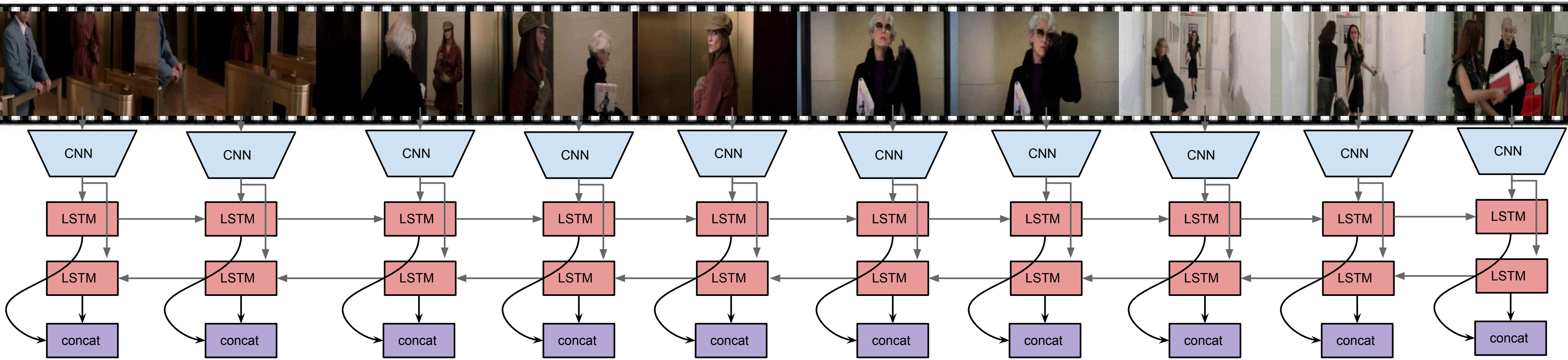
Unsupervised Temporal Segmentation

- Unsupervised method to identify change points
  - Kernel Temporal Segmentation [1]
- Use CNN features

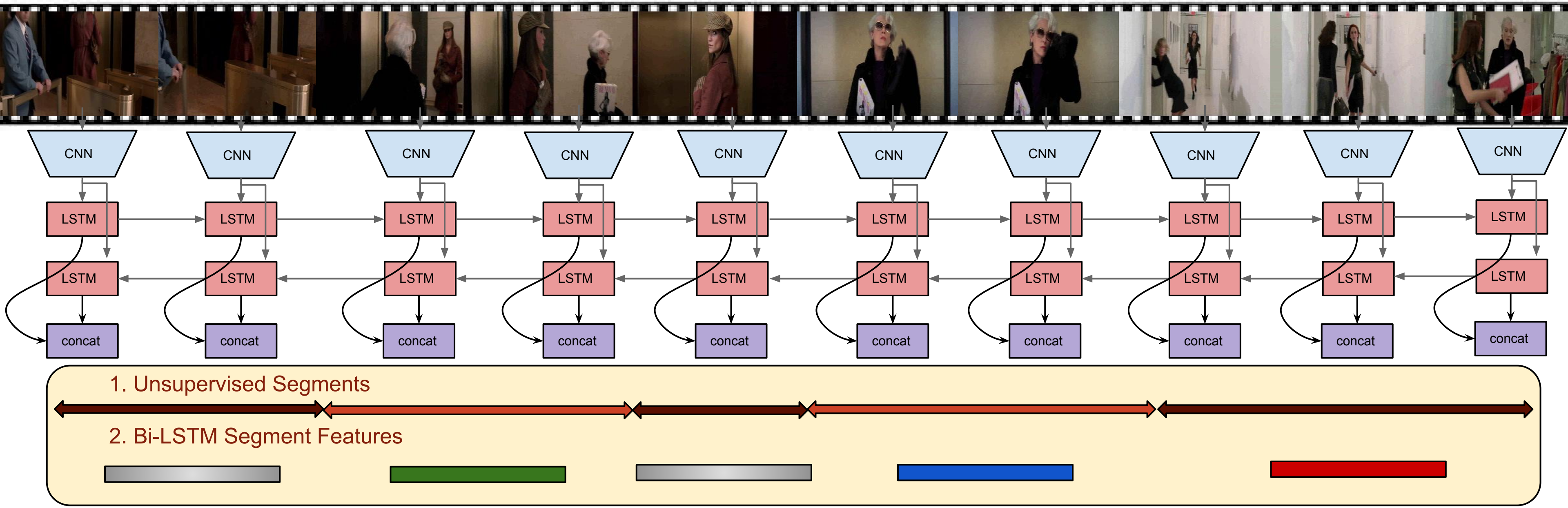
Unsupervised Coherent Segments

[1] D. Potapov, M. Douze, Z. Harchaoui, C. Schmid. ECCV'14

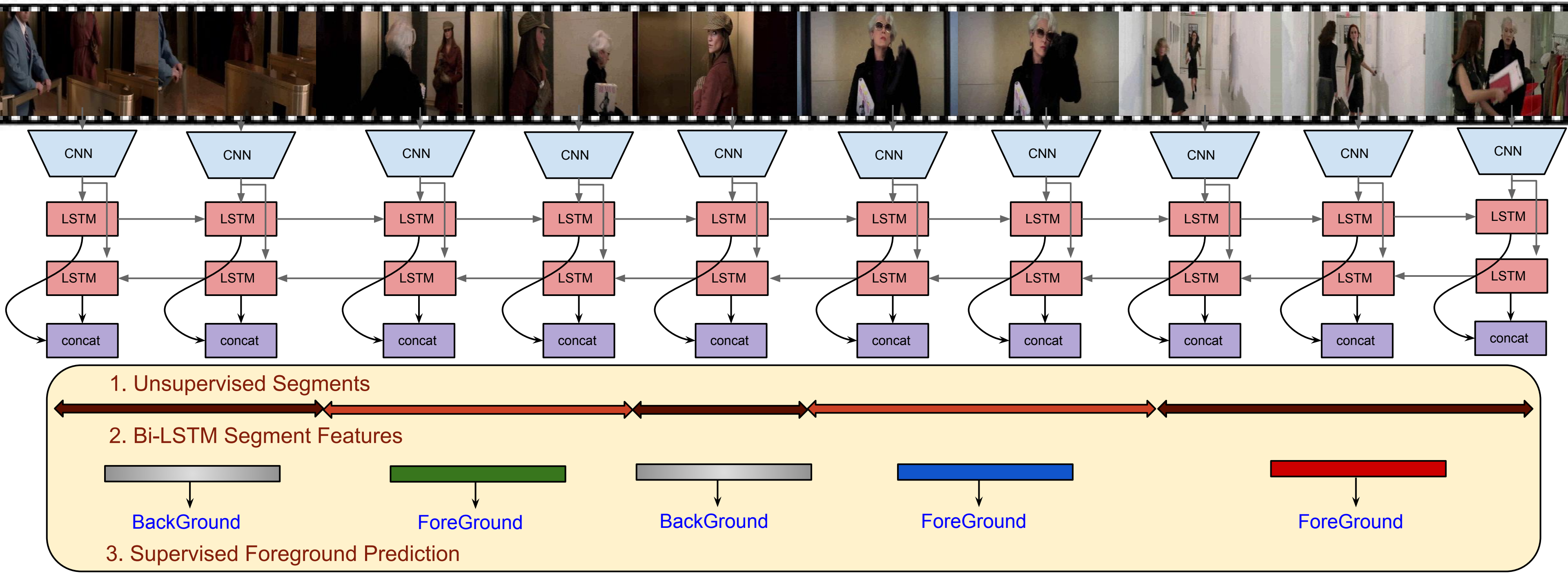
Bi-Directional LSTM encoder



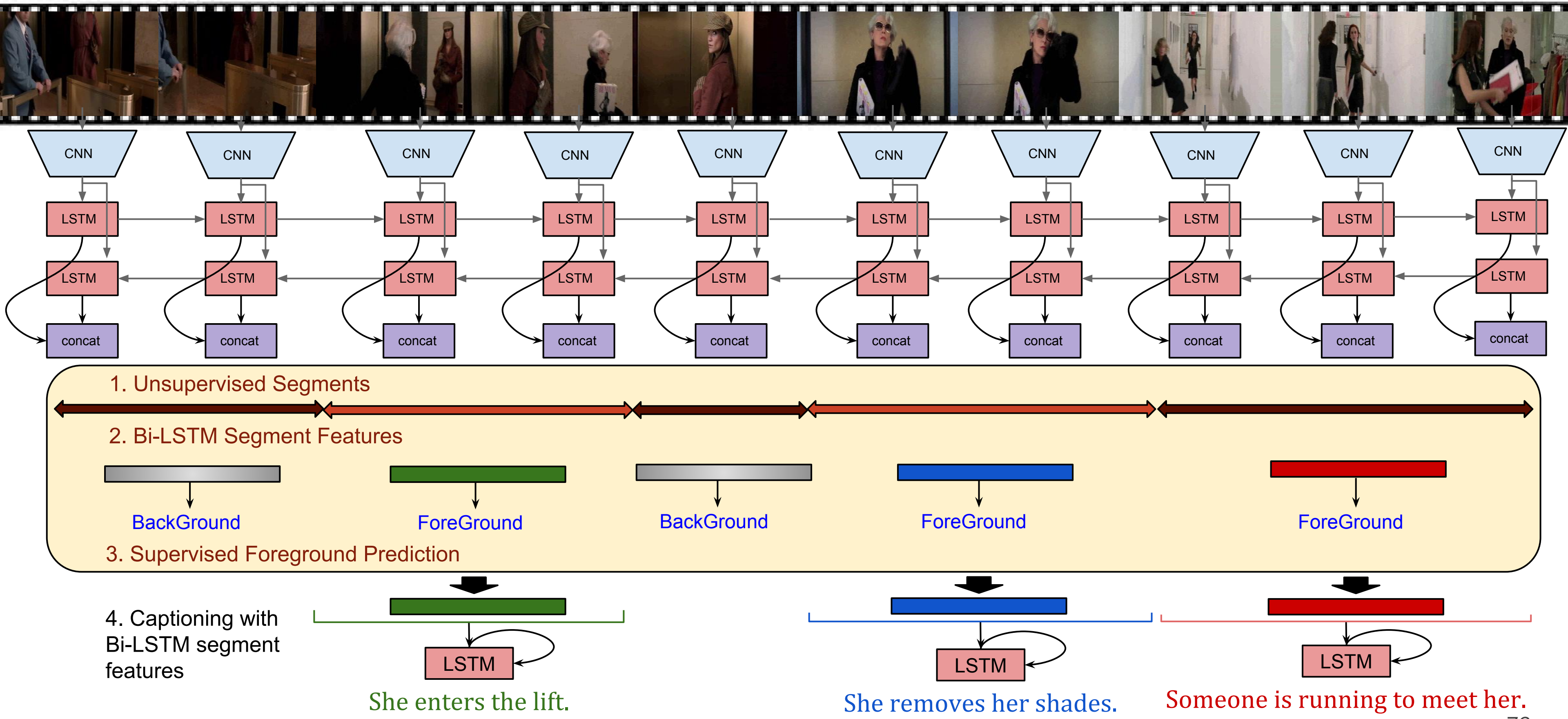
Segment Features



Supervised Foreground Prediction



Temporal Segmentation and Description (TSDN) Model

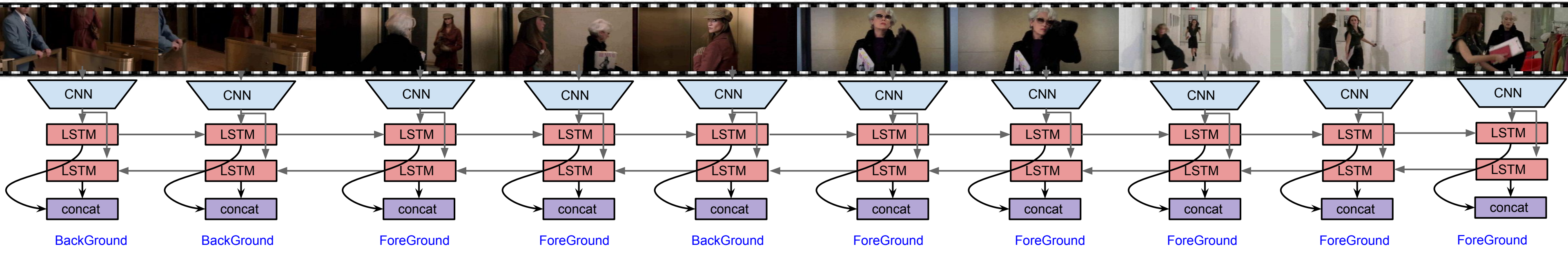


Models for Comparison

- Uniform Segments
- Scene Subshot
  - >= 40% change in pixel intensities between subsequent frames [Richardson ‘04]
- Kernel Temporal Segmentation (KTS)
  - All segments are foreground [Potapov et al. ECCV’14]
- Frame-wise foreground/background (FGBG)

Models for Comparison

- Uniform Segments
- Scene Subshot
    - >= 40% change in pixel intensities between subsequent frames
- Kernel Temporal Segmentation
    - All segments are foreground
- Frame-wise foreground/background (FGBG)



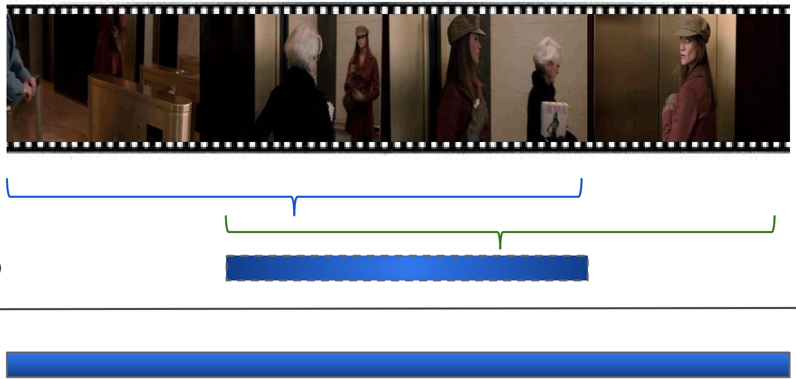
Datasets

- Full-length movies cut to ~1min clips
- Non-overlapping segments

Our Dataset	MPII-MD	M-VAD
Num. of movies	94	92
Num. of clips	11,560	8,789
Avg. clip length	57s	58s
Avg. num. segments per clip	6	6
Total Duration of clips	184h 46m	141h 42m
Num. segments/descriptions	68,375	56,431

Metrics - Segmentation

- F1 @ IoU threshold >= 0.5



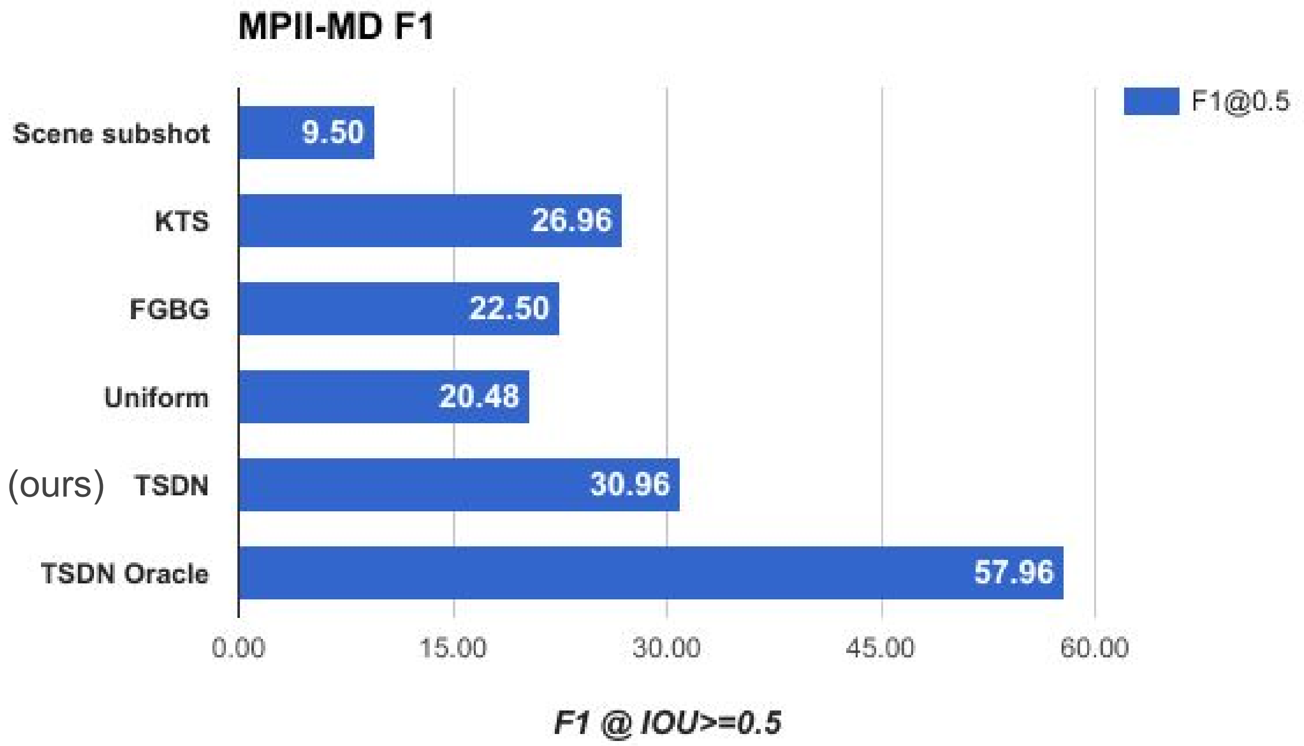
$$\text{IoU} = \frac{\text{Duration of overlap}}{\text{Duration of union}}$$

- For each groundtruth, pick distinct prediction with highest IoU.

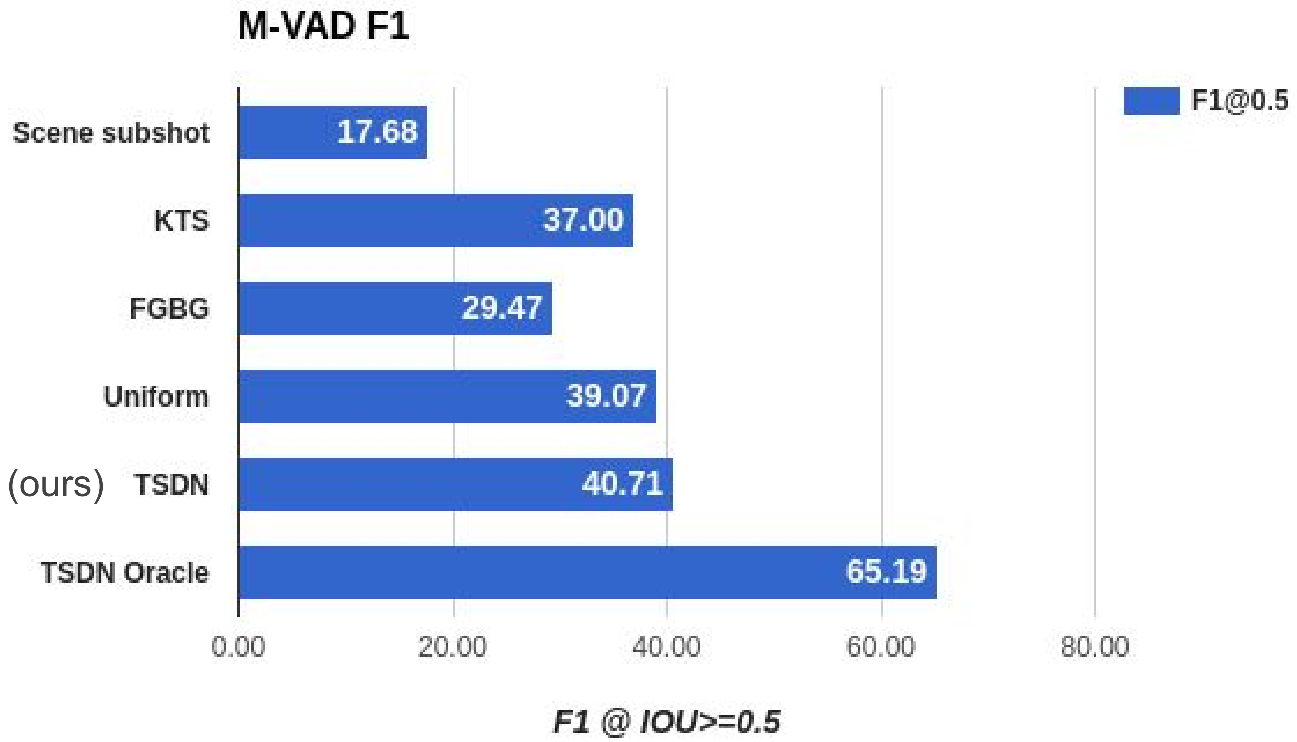
Metrics - Captioning

- METEOR (automated metric)
- Caption of 1 best segment with highest overlap

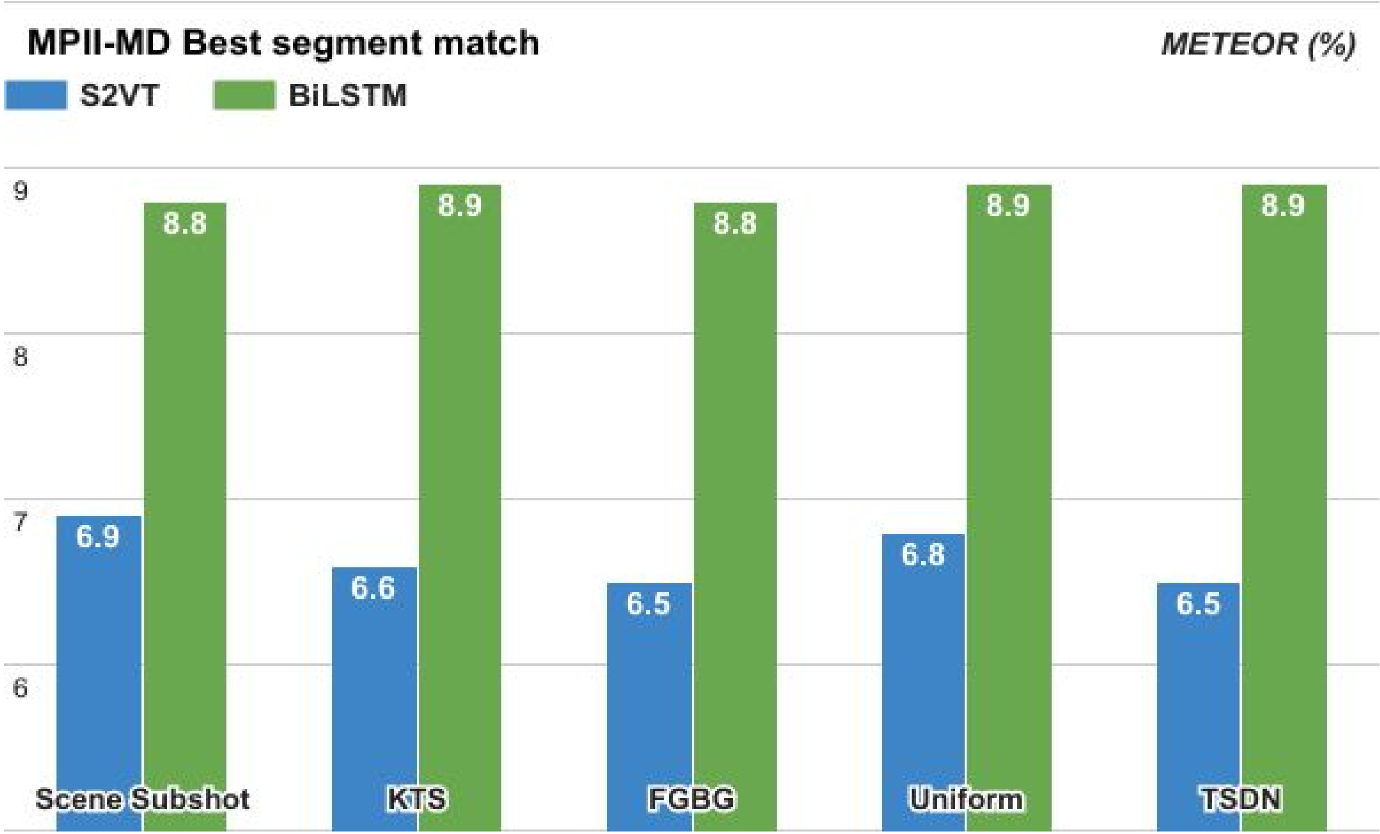
Segmentation : MPII-MD



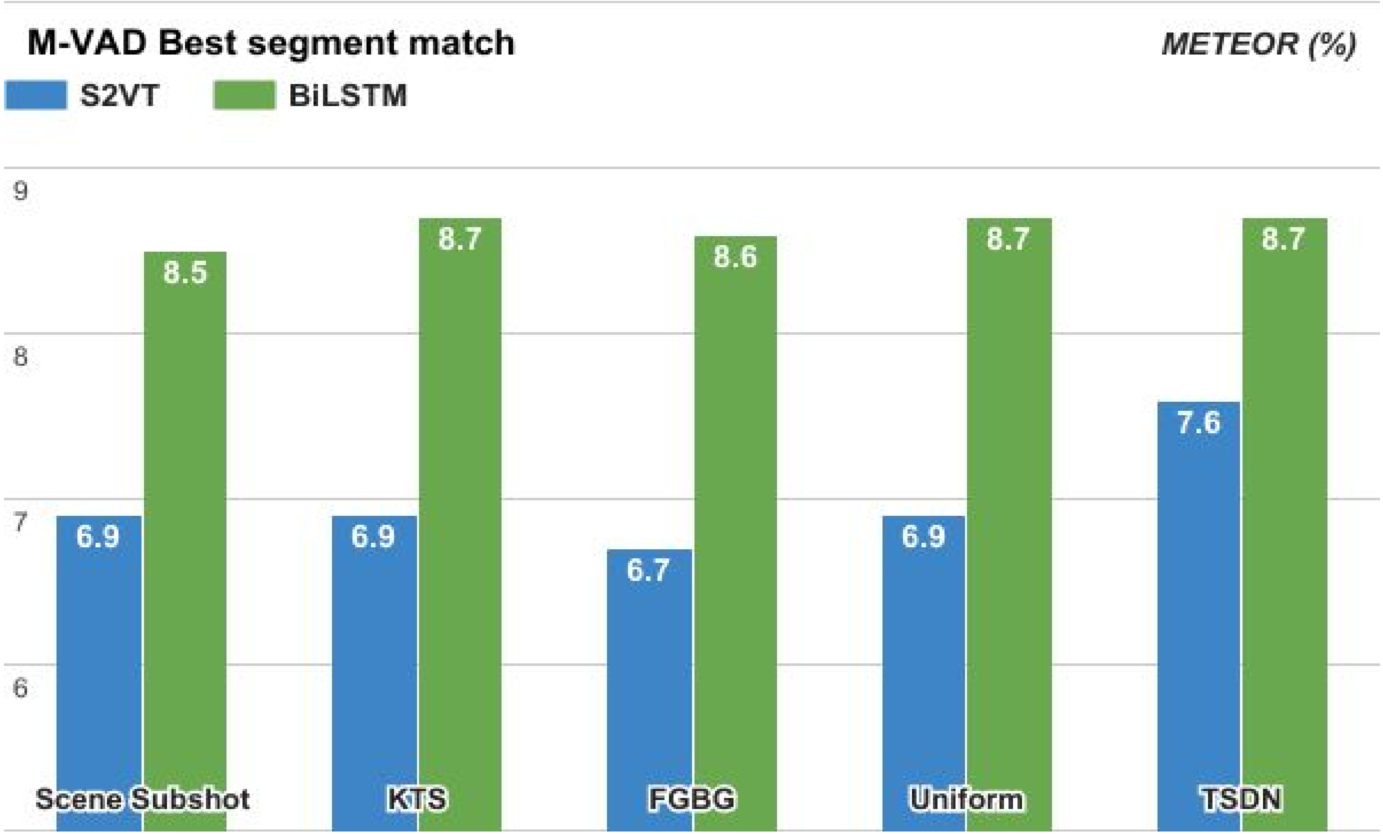
Segmentation : M-VAD



Captioning : MPII-MD (1 best)



Captioning : M-VAD (1 best)



Examples

Someone 's eyes widen.

Someone steps out of the room and shuts the door.

Someone opens the door and finds a photo of someone's name on the table.

Now, the sun shines on the horizon.

The car drives off the road and parks.

GT:

He hits the disconnect button.

Now, in someone's pink-tiled bathroom, someone searches a vanity then picks through dirty laundry strewn around the tub.

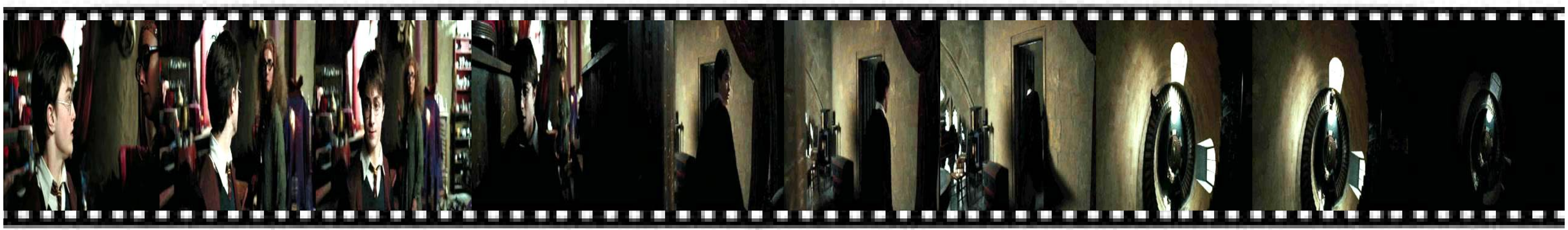
She finds a bar coaster in a pair of jeans.

Now, on her cell, she crosses the Verrazano-Narrows Bridge.

Uniform

KTS

Examples



Someone looks at someone, who's standing in the doorway

Someone walks out of the room and finds someone

Someone walks into the room and finds a small metal grill

The shape moves down the stairs, and the lights go out.

GT: Bemused, someone gazes at someone.

A worried look on his face, he runs out of the room and hurries away down the circular staircase

Uniform:

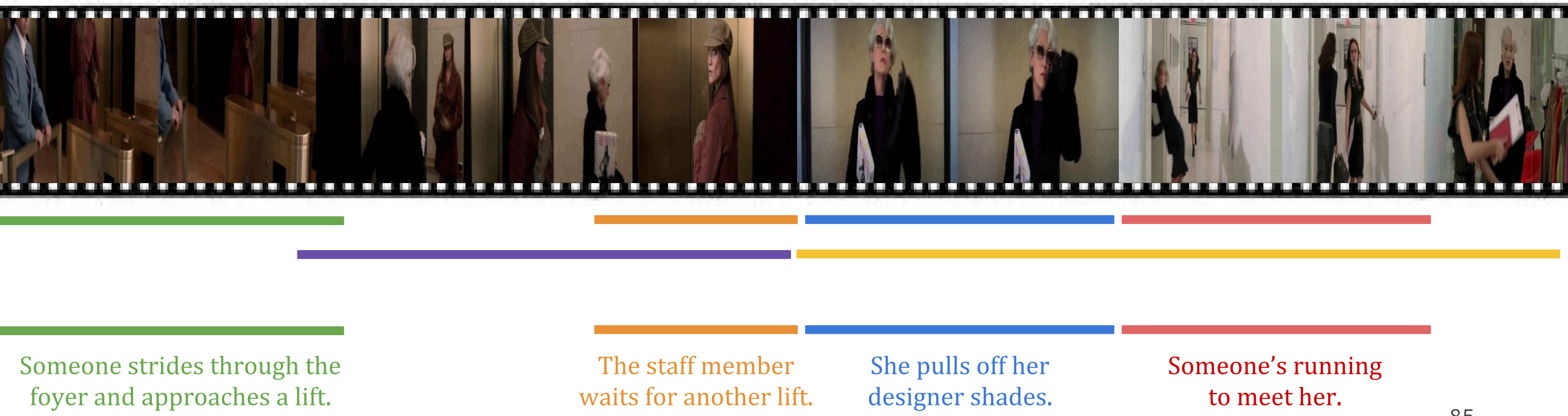
KTS:

Outline

- Review (proposal)
  - Background
  - Encoder-Decoder approaches to video description
- External knowledge to improve video description
- External knowledge for novel object captioning
- Temporal segmentation and description for long videos
- Future Directions

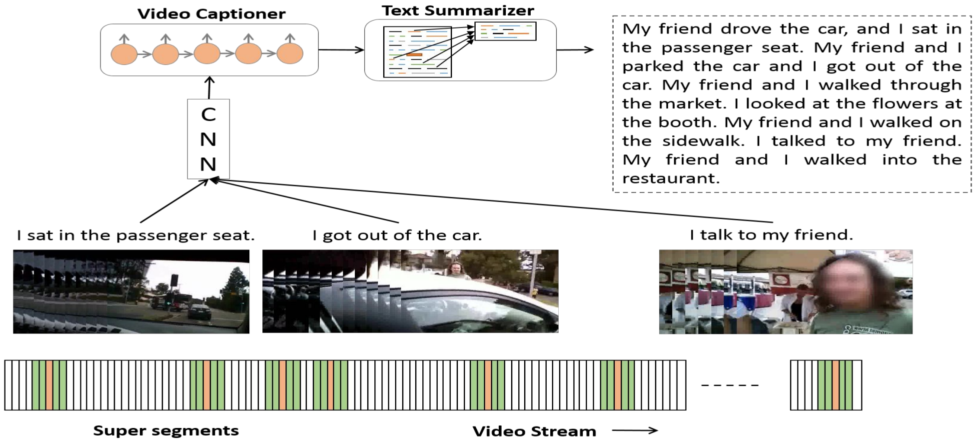
Future Directions

- Jointly segmenting and describing
  - Network to generate segment proposals



Future Directions

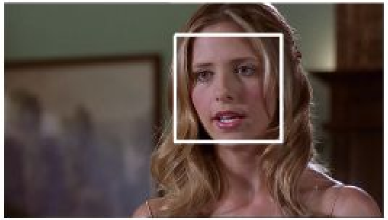
- Jointly segmenting and describing
  - Network to generate segment proposals
- Textual summarization of videos
  - Ego-centric videos



S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prudhommeaux, R. Ptucha. WACV ‘17

Future Directions

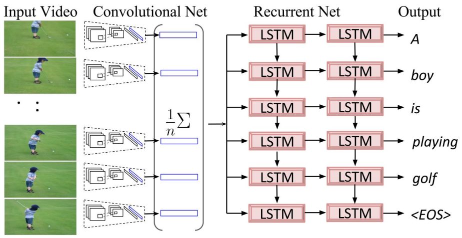
- Jointly segmenting and describing
    - Network to generate segment proposals
- Textual summarization of videos
    - Ego-centric videos
- Fully automating DVS for movies
    - Multimodal captioning (+audio) [Ramanishka et al. ACMMM’16]
    - Handling names of characters/actors

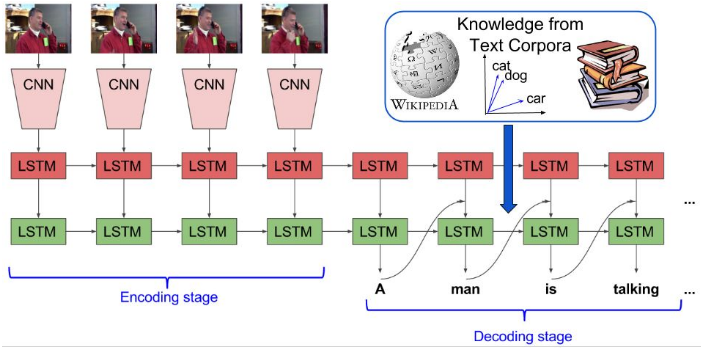


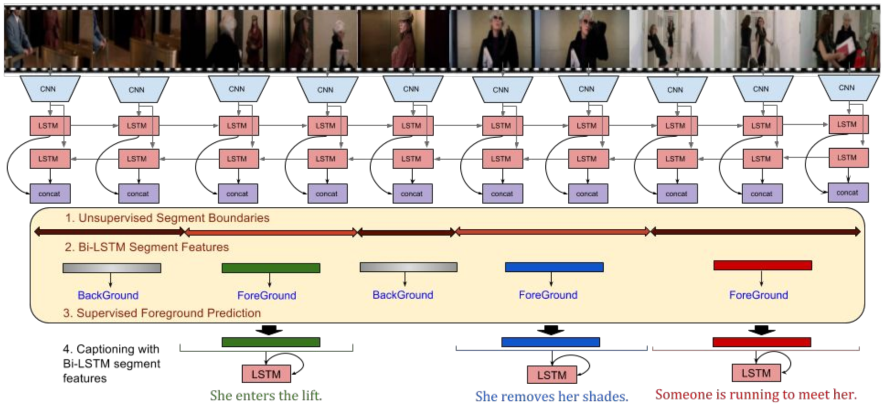

Future Directions

- Jointly segmenting and describing
  - Network to generate segment proposals
- Textual summarization of videos
  - Ego-centric videos
- Fully automating DVS for movies
  - Multimodal captioning (+audio)
  - Handling names of characters/actors

Conclusion







1. Deep architectures for video description.
2. Jointly models a sequence of frames and sequence of words.
3. Incorporating external linguistic knowledge.
4. Describe novel objects.
5. Temporally segment and describe long videos.

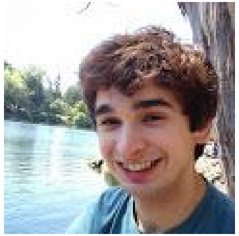
Evaluation on Youtube videos and movie corpora.

Collaborators

Raymond
Mooney

Trevor
Darrell

Kate
Saenko



Jeff
Donahue

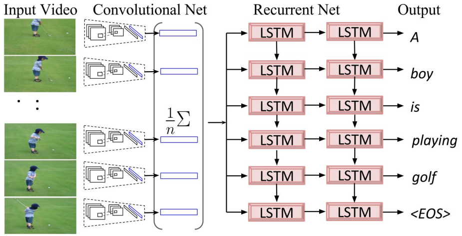
Marcus
Rohrbach

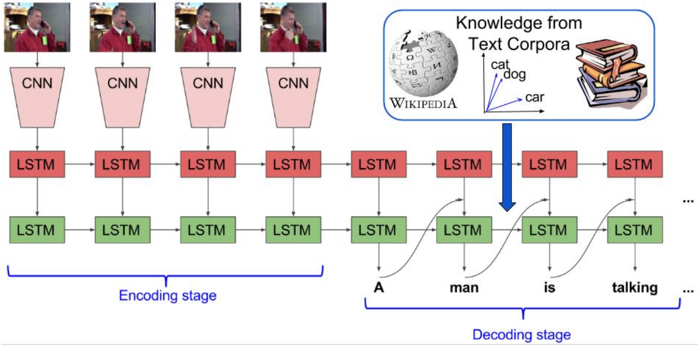
Lisa Anne
Hendricks

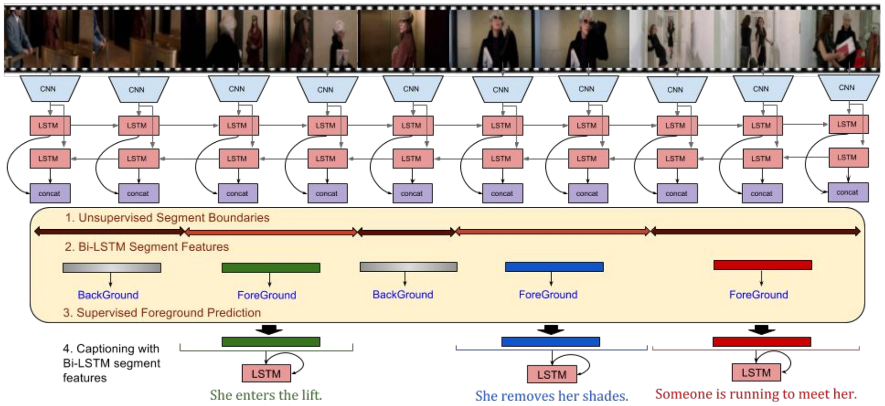
Huijuan
Xu

Vasili
Ramanishka

Thanks!







1. Deep architectures for video description.
2. Jointly models a sequence of frames and sequence of words.
3. Incorporating external linguistic knowledge.
4. Describe novel objects.
5. Temporally segment and describe long videos.

Evaluation on Youtube videos and movie corpora.

Project Pages and Code for models

Mean-pool:

https://vsubhashini.github.io/naacl15_project.html

S2VT:

http://vsubhashini.github.io/s2vt.html#code

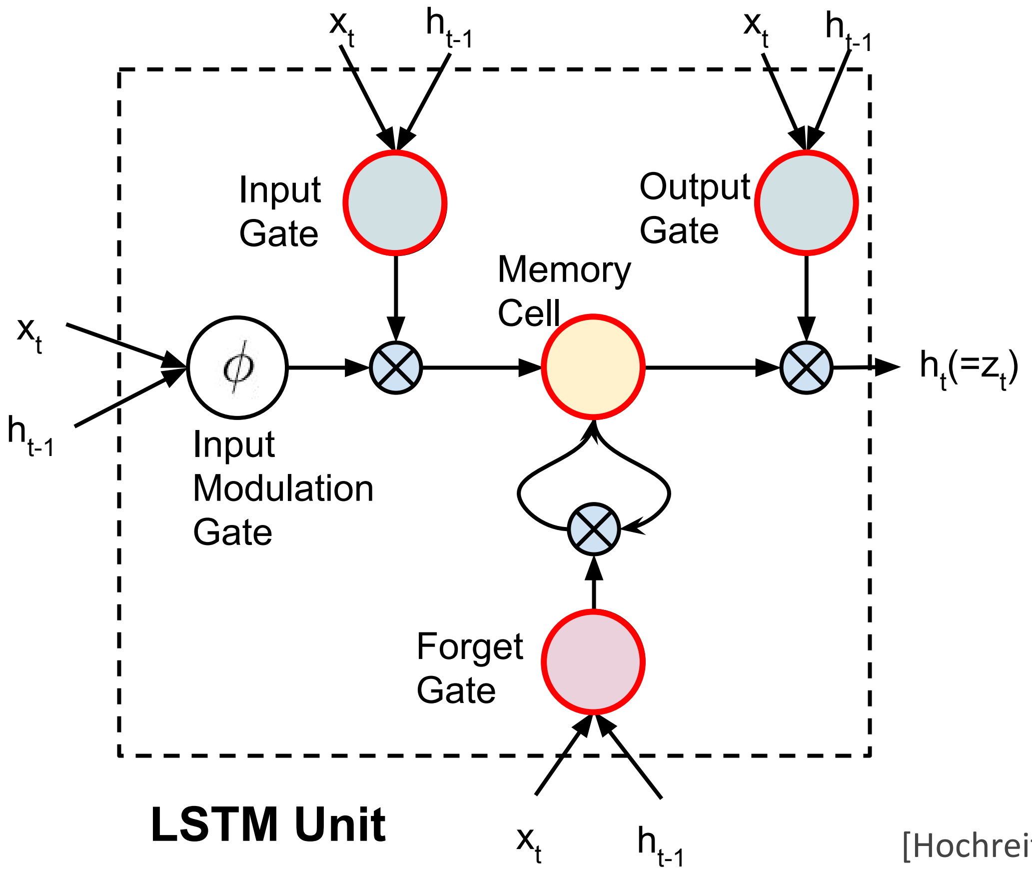
Incorporating linguistic knowledge:

http://vsubhashini.github.io/language_fusion.html

Novel Object Captioning:

http://vsubhashini.github.io/noc.html

[Background] LSTM



$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1})$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1})$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi(W_{xc}x_t + W_{hc}h_{t-1})$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1})$$

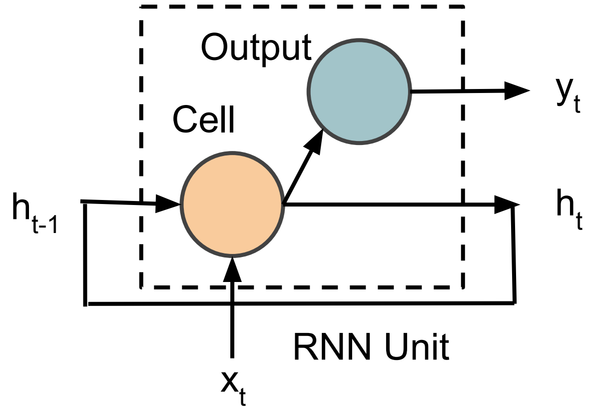
$$h_t = o_t \odot \phi(c_t)$$

[Hochreiter and Schmidhuber '97]
[Graves '13]

Recurrent Neural Networks

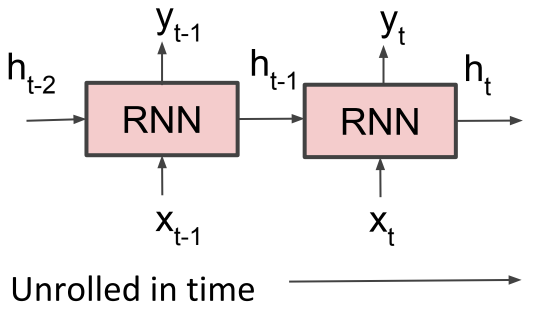
Successful in translation, speech.

RNNs can map an input to an output sequence.



$$h_t = \sigma(W^{hh} h_{t-1} + W^{hx} x_t)$$

Pr(out y_t | input, out $y_0 \ldots y_{t-1}$)



Problems:

1. Hard to capture long term dependencies
2. Vanishing gradients (shrink through many layers)

One Solution: Long Short Term Memory (LSTM) unit