Video Captioning

Erin Grant

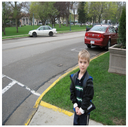
March 1st, 2016

Last Class: Image Captioning

there is a cat
sitting on a shelf .

a plate with a fork
and a piece of cake .

a black and white
photo of a window .



a young boy standing
on a parking lot
next to cars .

a wooden table
and chairs arranged
in a room .

a kitchen with
stainless steel
appliances .

this is a herd
of cattle out
in the field .

a car is parked
in the middle
of nowhere .

a ferry boat on
a marina with a
group of people .

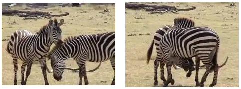
a little boy with
a bunch of friends
on the street .

From Kiros et al. [2014]

This Week: Video Captioning

AKA: Image captioning through time!

S2VT: A man is doing stunts on his bike.



S2VT: A herd of zebras are walking in a field.

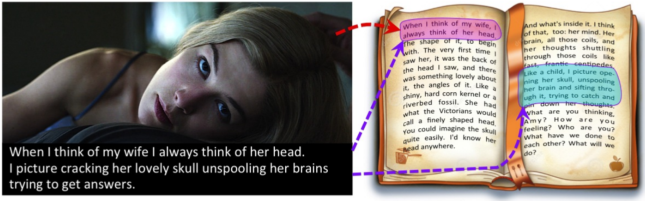
S2VT: A young woman is doing her hair.

S2VT: A man is shooting a gun at a target.

From Venugopalan et al. [2015]

Related Work (1)

Toronto: Joint Embedding from **Skip-thoughts + CNN** :



When I think of my wife I always think of her head.
I picture cracking her lovely skull unspooling her brains
trying to get answers.

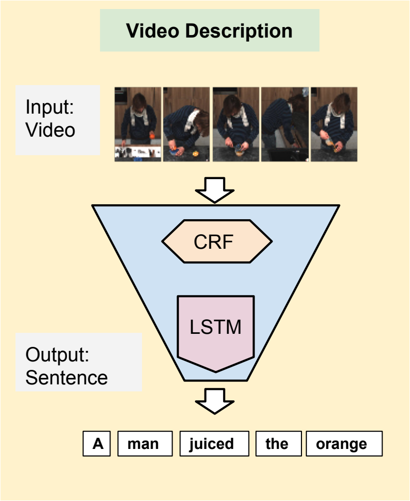
When I think of my wife, I always think of her head. The shape of it, to begin with. The very first time I saw her, it was the back of the head I saw, and there was something lovely about it, the angles of it. Like a shiny, hard corn kernel or a riverbed fossil. She had what the Victorians would call a finely shaped head. You could imagine the skull quite easily. I'd know her head anywhere.

And what's inside it. I think of that, too: her mind. Her brain, all those coils, and her thoughts shuttling through those coils like fast, frantic centipedes. Like a child, I picture opening her skull, unspooling her brain and sifting through it, trying to catch and pin down her thoughts. What are you thinking, Amy? How are you feeling? Who are you? What have we done to each other? What will we do?

from Zhu et al. [2015]: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books

Related Work (2)

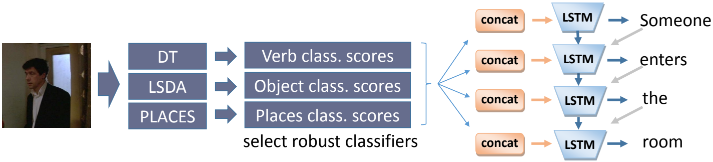
Berkeley: **Long-term Recurrent Convolutional Networks**:



From Donahue et al. [2015]: Long-term recurrent convolutional networks for visual recognition and description

Related Work (3)

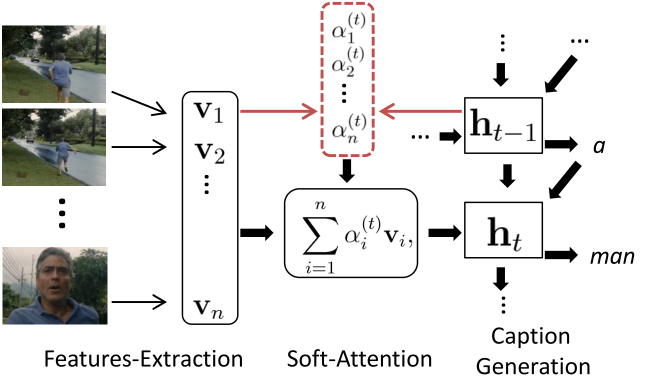
MPI: Ensemble of **weak classifiers** + **LSTM**:



from Rohrbach et al. [2015]: The long-short story of movie description

Related Work (4)

Montréal: **(SIFT, HOG) Features + 3-D CNN + LSTM + Attention**:



From Yao et al. [2015]: Video description generation incorporating spatio-temporal features and a soft-attention mechanism

We can simplify the problem. . .

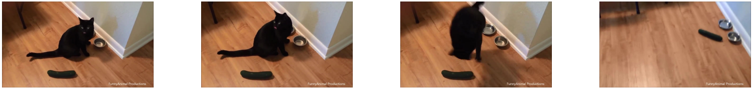
In captioning, we translate one modality **(image)** to another **(text)**.

Image captioning : **Fixed** length sequence (image) to **variable** length sequence (words).

Video captioning : **Variable** length sequence (video frames) to **variable** length sequence (words).

Formulation

- Let $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be the sequence of **video frames**.



- Let $(y_1, \ldots, y_m)$ be the sequence of **words**.

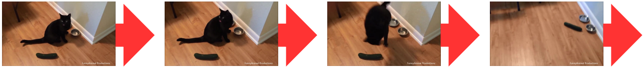
  (The, cat, is, afraid, of, the, cucumber.)

- We want to maximise $p(y_1, \ldots, y_m \mid \mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Formulation contd.

Idea:

- Accumulate the sequence of video frames into a single **encoded** vector.

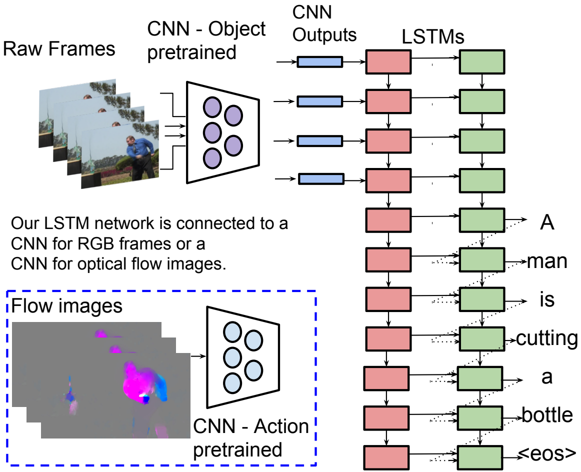


$$\begin{bmatrix}126\\42\\\vdots\\429\\12\end{bmatrix}$$

- **Decode** that vector into words one-by-one.

$\begin{bmatrix}126\\42\\\vdots\\429\\12\end{bmatrix}$ → The $\Longrightarrow$ cat $\Longrightarrow$ is $\Longrightarrow$ afraid $\Longrightarrow$ of . . .

The S2VT Model



Our LSTM network is connected to a CNN for RGB frames or a CNN for optical flow images.

From Venugopalan et al. [2015]: Sequence to sequence-video to text

Optimization

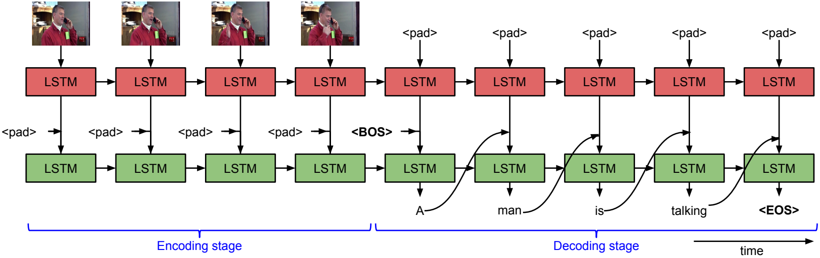
During decoding, maximise

$$\log p(y_1, \ldots, y_m \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) \\ = \sum_{t=1}^{m} \log p(y_t \mid h_{n+1-1}, y_{t-1}))$$

Train using stochastic gradient descent.

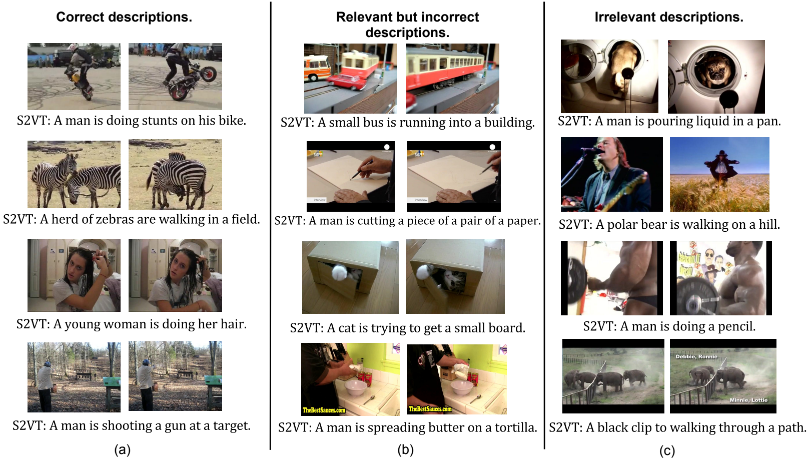
Encoder weights are jointly updated with decoder weights because we are backpropagating through time.

S2VT Model in Detail



From Venugopalan et al. [2015]

S2VT Results (Qualitative)



Correct descriptions.

S2VT: A man is doing stunts on his bike.

S2VT: A herd of zebras are walking in a field.

S2VT: A young woman is doing her hair.

S2VT: A man is shooting a gun at a target.

(a)

Relevant but incorrect descriptions.

S2VT: A small bus is running into a building.

S2VT: A man is cutting a piece of a pair of a paper.

S2VT: A cat is trying to get a small board.

S2VT: A man is spreading butter on a tortilla.

(b)

Irrelevant descriptions.

S2VT: A man is pouring liquid in a pan.

S2VT: A polar bear is walking on a hill.

S2VT: A man is doing a pencil.

S2VT: A black clip to walking through a path.

(c)

From Venugopalan et al. [2015]

S2VT Results (Quantitative)

Model	METEOR
FGM Thomason et al. [2014]	23.9
Mean pool	
- AlexNet Venugopalan et al. [2015]	26.9
- VGG	27.7
- AlexNet COCO pre-trained Venugopalan et al. [2015]	29.1
- GoogleNet Yao et al. [2015]	28.7
Temporal attention	
- GoogleNet Yao et al. [2015]	29.0
- GoogleNet + 3D-CNN Yao et al. [2015]	29.6
S2VT	
- Flow (AlexNet)	24.3
- RGB (AlexNet)	27.9
- RGB (VGG) random frame order	28.2
- RGB (VGG)	29.2
- RGB (VGG) + Flow (AlexNet)	**29.8**

Table: Microsoft Video Description (MSVD) dataset (METEOR in %, higher is better).

From Venugopalan et al. [2015]

Datasets

- Microsoft Video Description corpus (MSVD) Chen and Dolan [2011]
    - web clips with human-annotated sentences
- MPII Movie Description Corpus (MPII-MD) Rohrbach et al. [2015] and Montreal Video Annotation Dataset (M-VAD) Yao et al. [2015]
    - movie clips with captions sourced from audio/script

Resources

- Implementation of S2VT: Sequence-to-Sequence Video-to-Text
- Microsoft Video Description corpus (MSVD)
- MPII Movie Description Corpus (MPII-MD)
- Montreal Video Annotation Dataset (M-VAD)

References I

D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.

J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

References II

A. Rohrbach, M. Rohrbach, and B. Schiele. The long-short story of movie description. In *Pattern Recognition*, pages 209–221. Springer, 2015.

J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, volume 2, page 9, 2014.

S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4534–4542, 2015.

L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Video description generation incorporating spatio-temporal features and a soft-attention mechanism. *arXiv preprint arXiv:1502.08029*, 2015.

References III

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27, 2015.