Deep Learning for Computer Vision

Vision and Language: Image Captioning

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

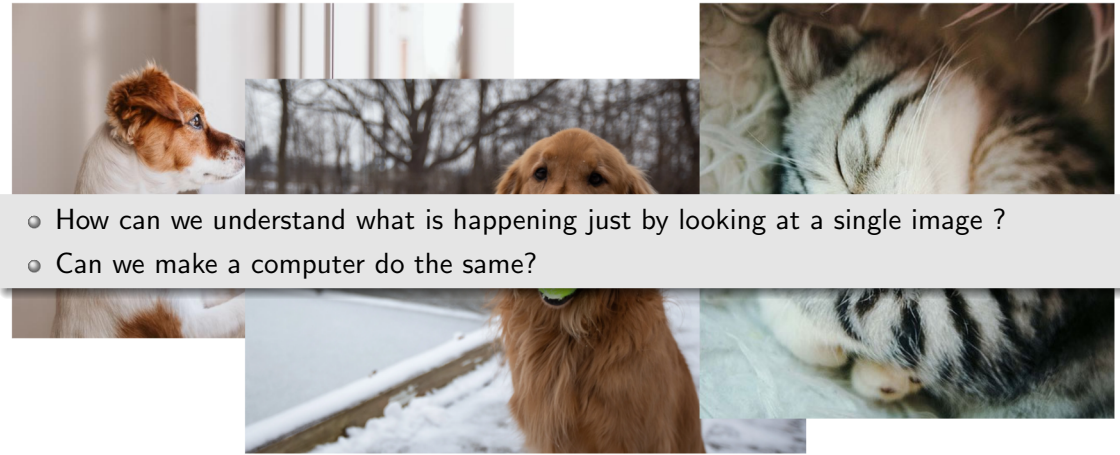
Review: Are autoencoders (AE) and PCA connected?

Review: Are autoencoders (AE) and PCA connected?

- Yes!

Describe these images

Describe these images



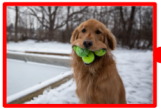
- How can we understand what is happening just by looking at a single image ?
- Can we make a computer do the same?

How to make a computer describe an image?

How to make a computer describe an image?

Some Method

How to make a computer describe an image?

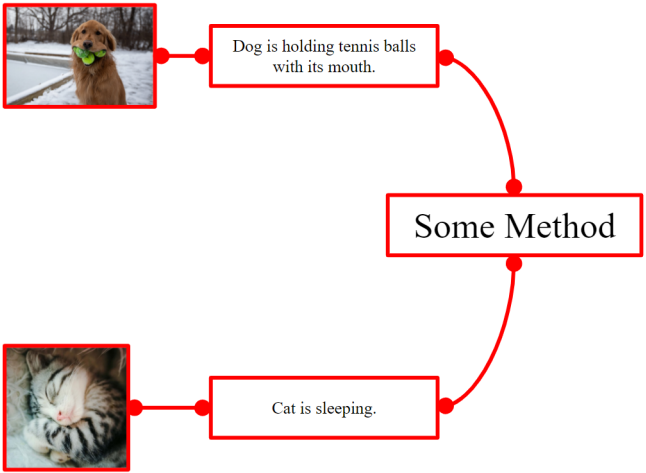


Dog is holding tennis balls with its mouth.

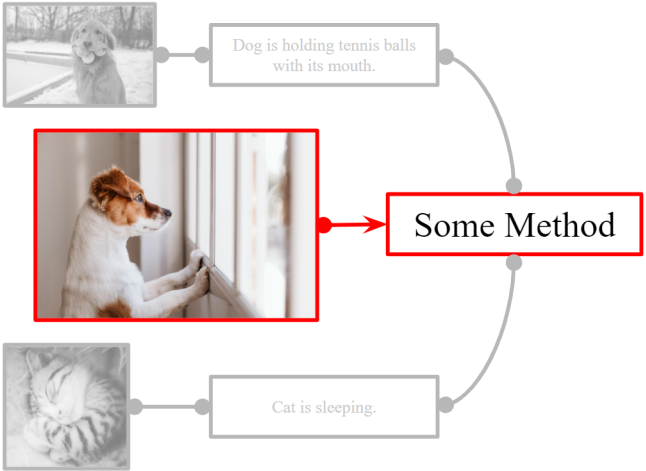
Some Method

Cat is sleeping.

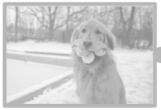
How to make a computer describe an image?



How to make a computer describe an image?



How to make a computer describe an image?



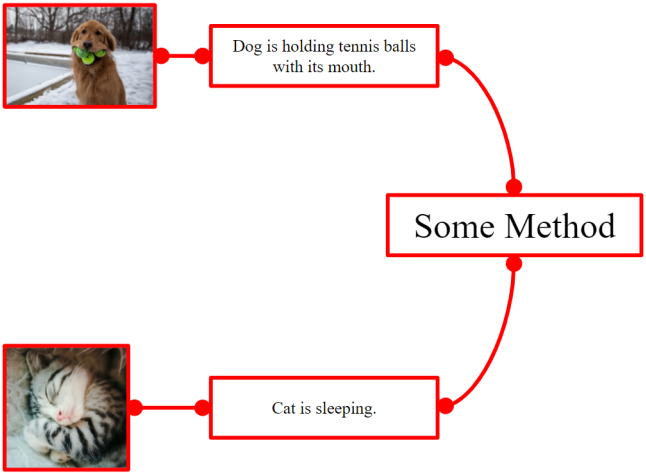
Dog is holding tennis balls with its mouth.

Some Method

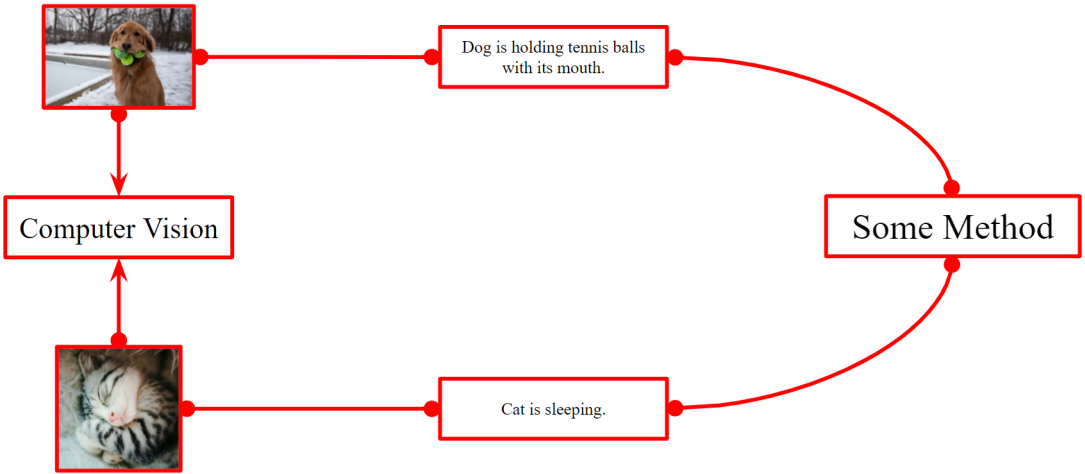
Dog is looking outside the window.

Cat is sleeping.

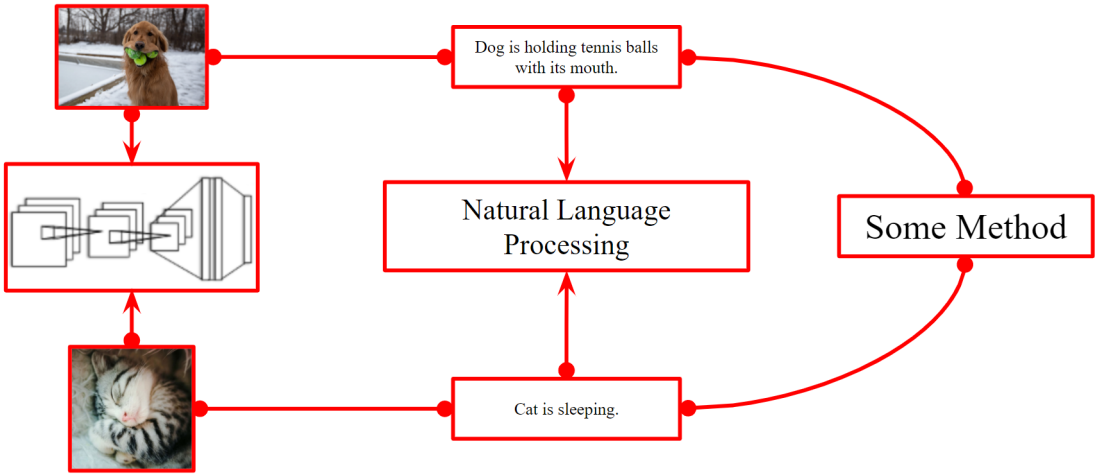
How to make a computer describe an image?



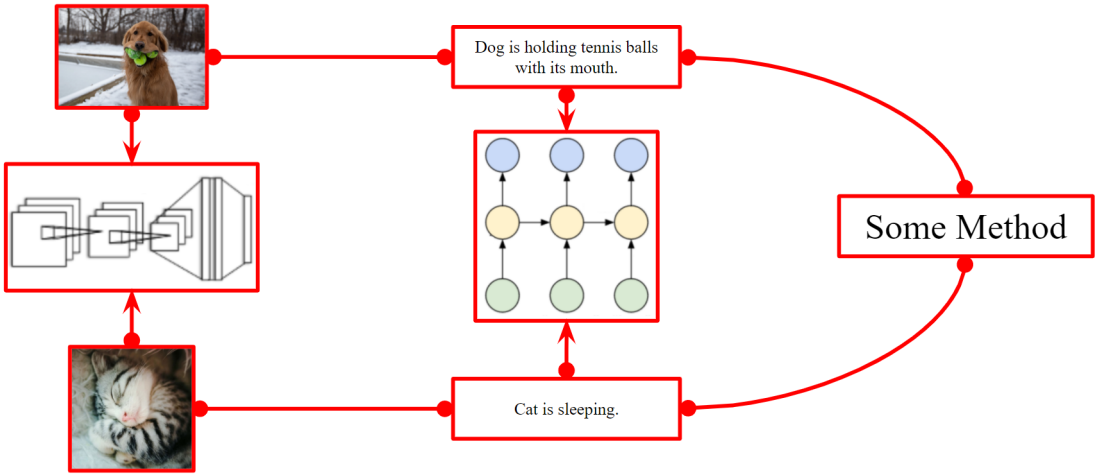
How to make a computer describe an image?



How to make a computer describe an image?



How to make a computer describe an image?



How to make a computer describe an image?

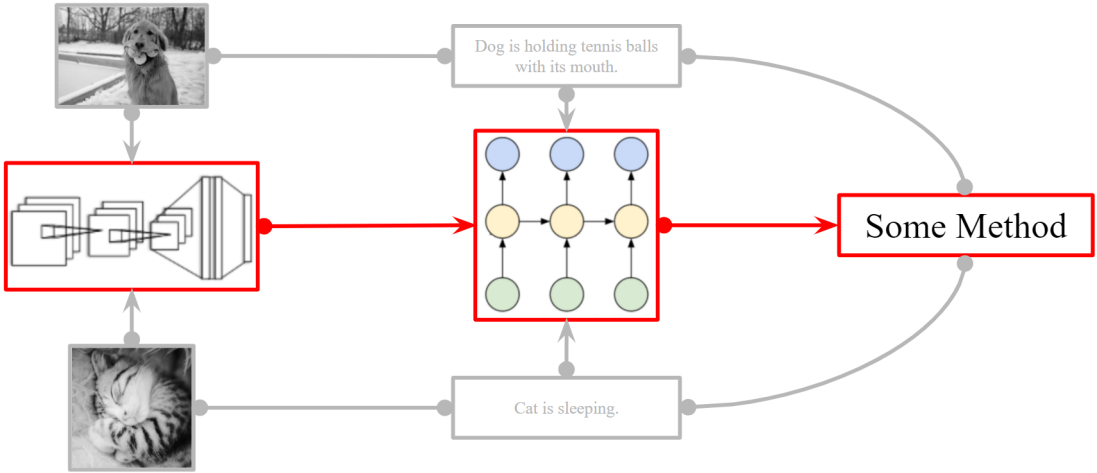
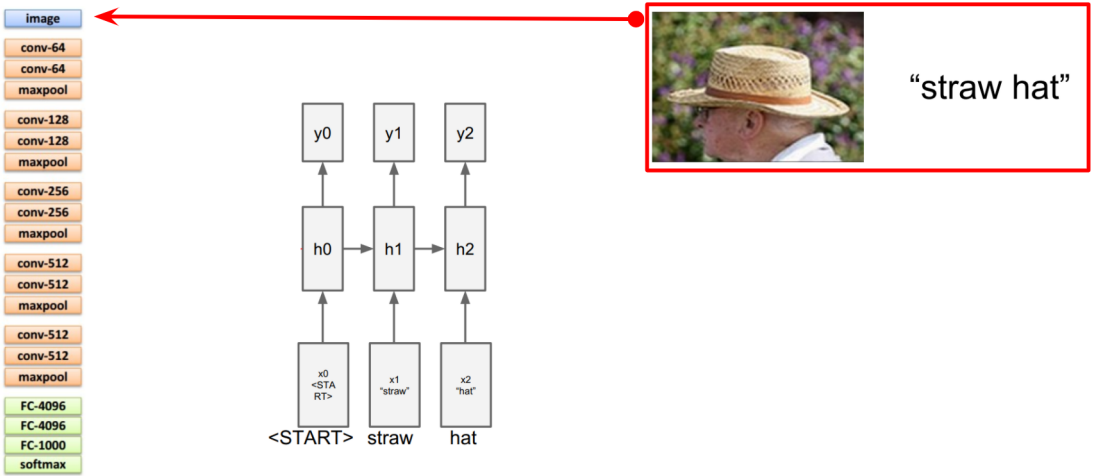


Image Captioning: Training

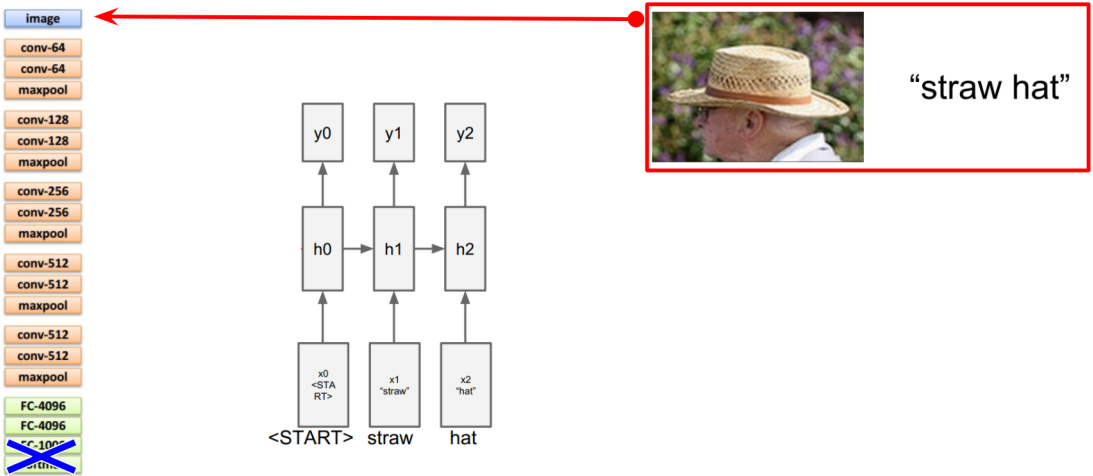
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Training



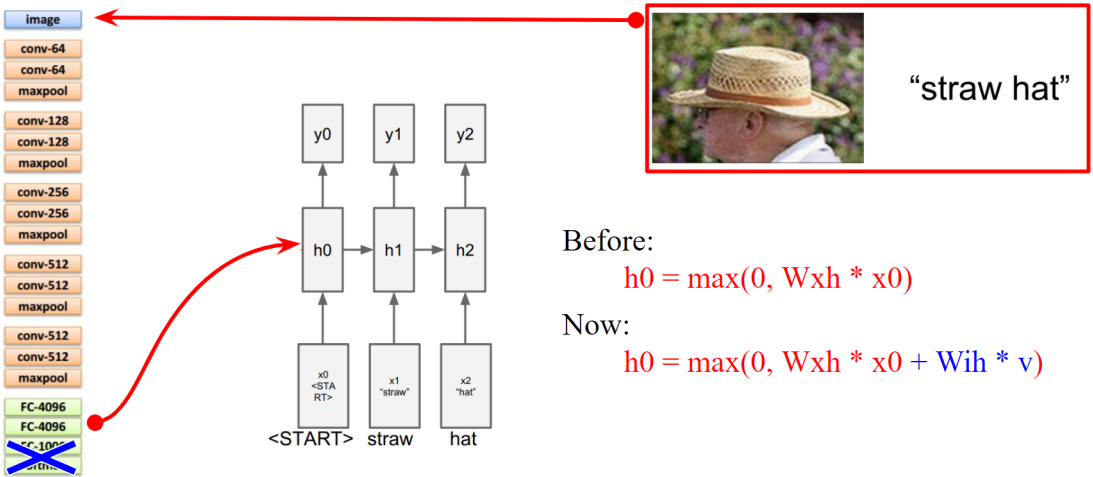
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Training



Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Training



Before:

$h0 = \max(0, Wxh * x0)$

Now:

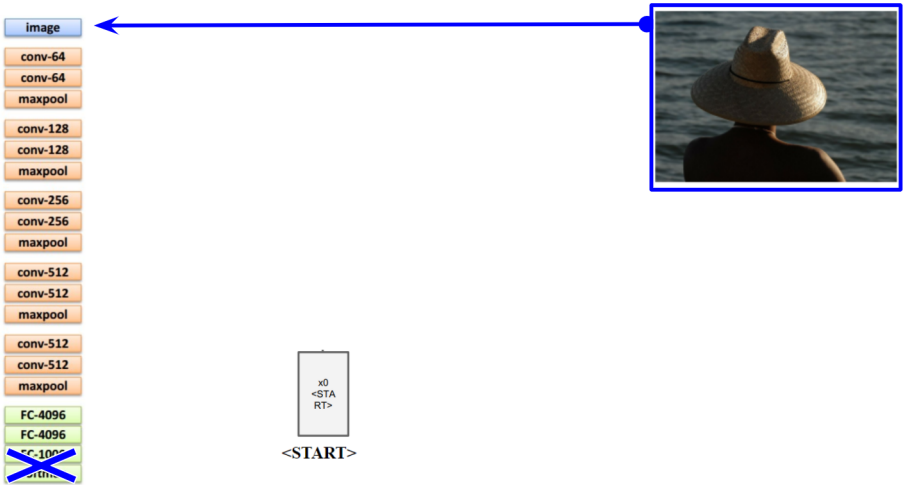
$h0 = \max(0, Wxh * x0 + Wih * v)$

Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)

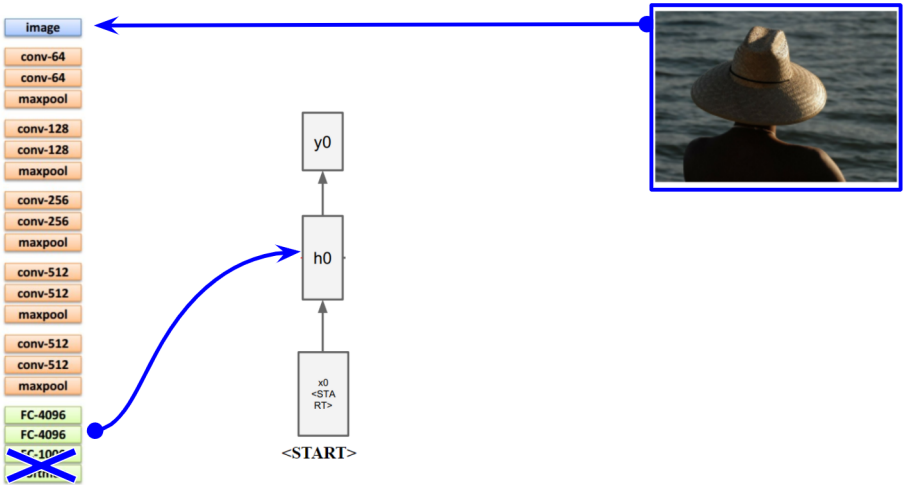
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



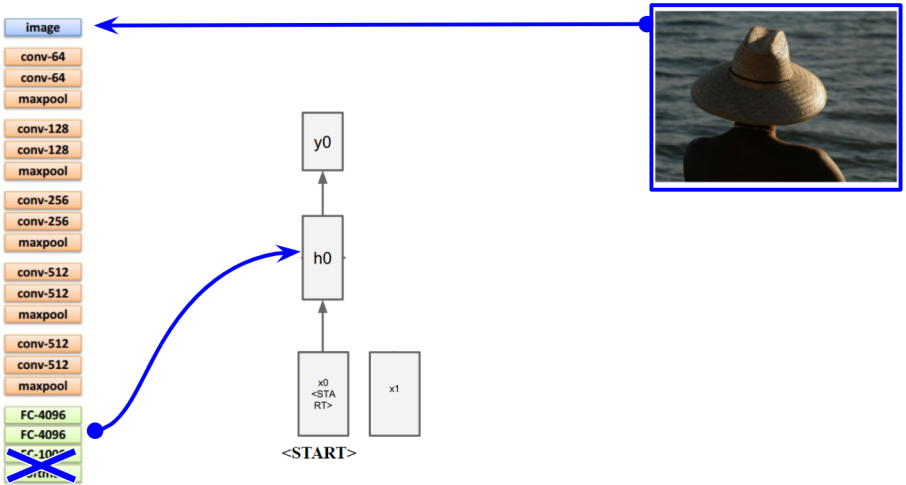
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



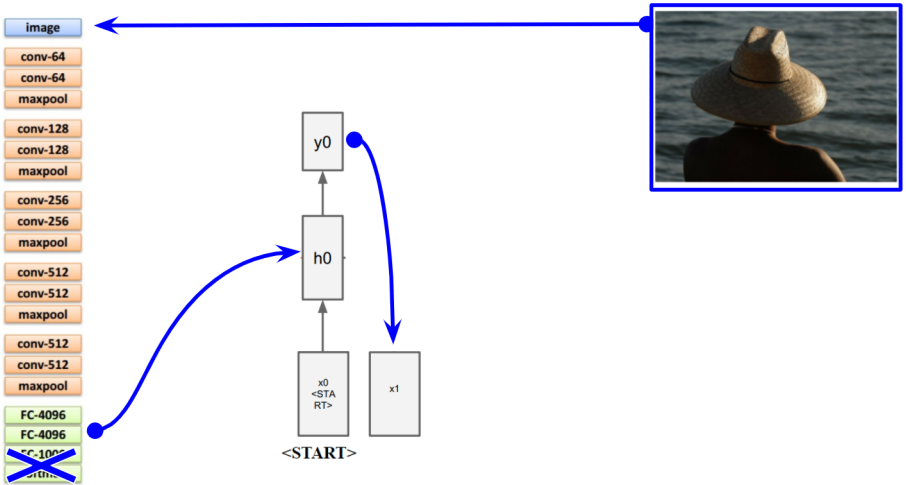
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



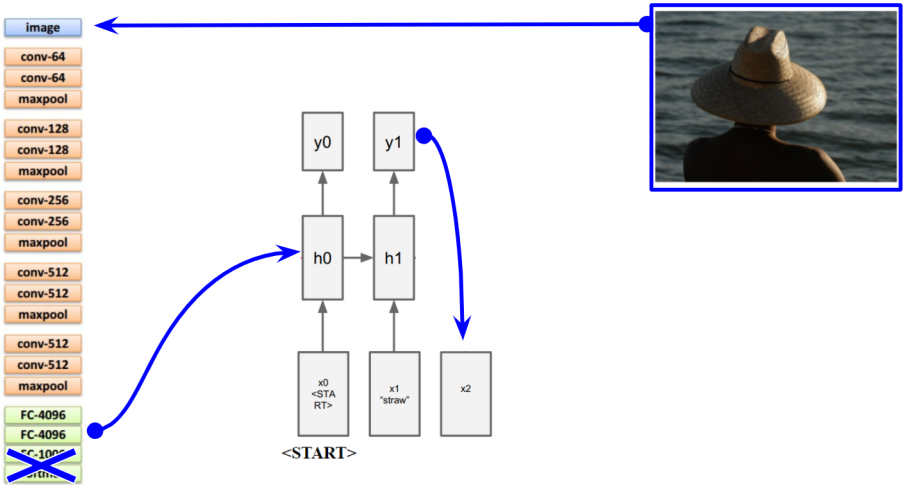
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



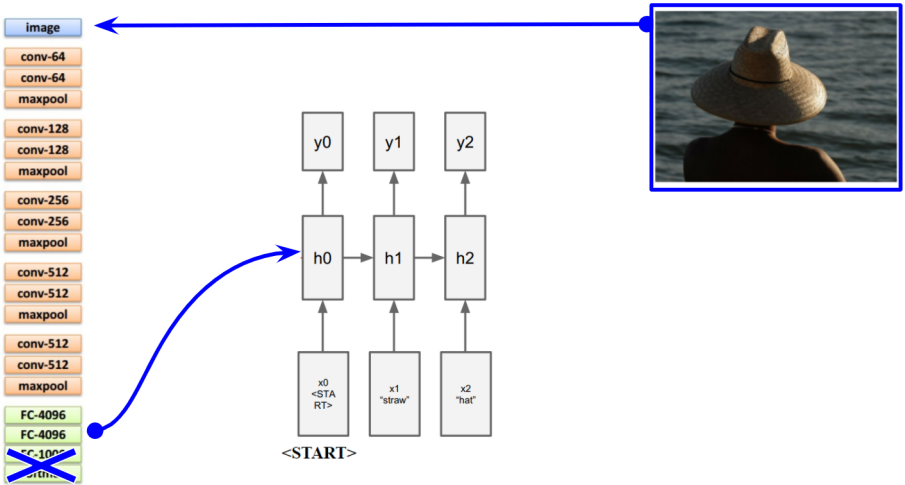
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



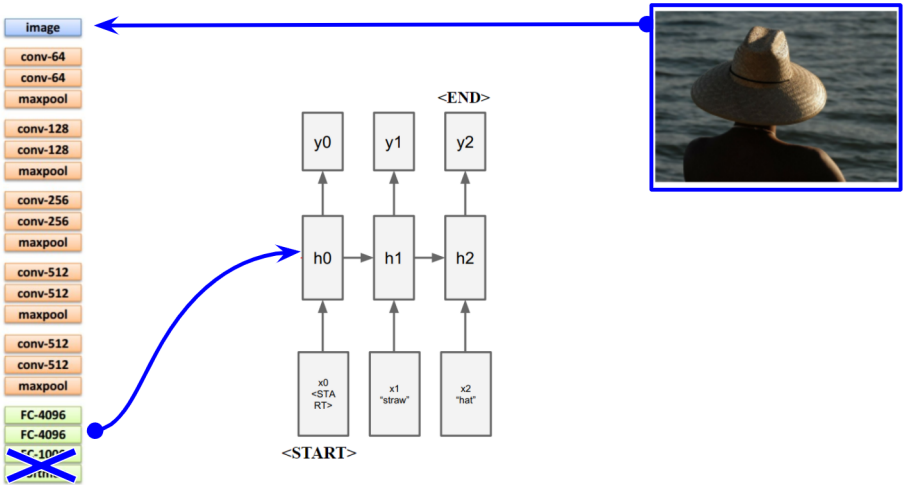
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



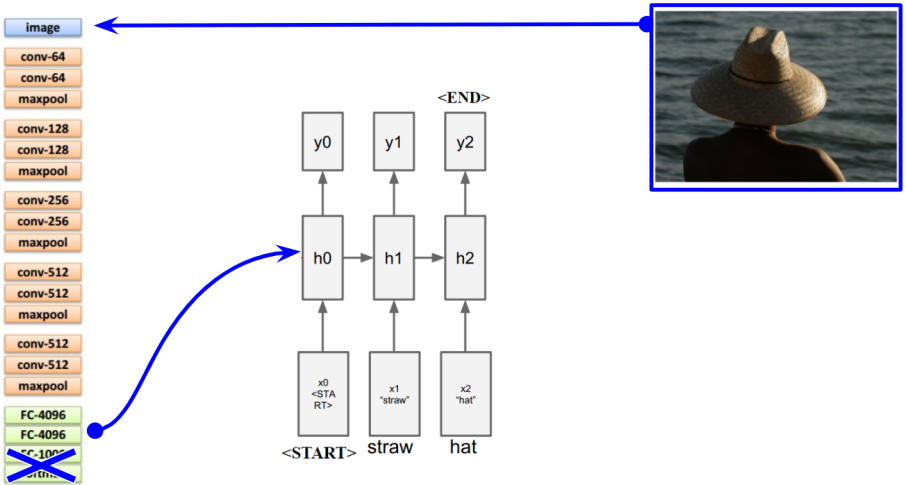
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

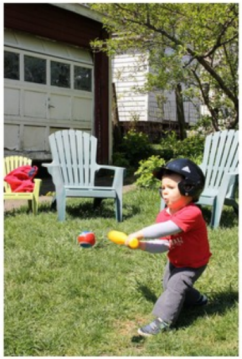
Image Captioning: Inference (Test Time)



Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Results

a group of people standing around a room with remotes
logprob: -9.17



a young boy is holding a baseball bat
logprob: -7.61

a cow is standing in the middle of a street
logprob: -8.84

Results: Failure Cases

Possible to understand why the method failed

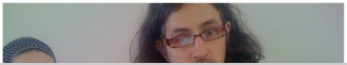
a man standing next to a clock on a wall
logprob: -10.08

a young boy is holding a baseball bat
logprob: -7.65

a cat is sitting on a couch with a remote control
logprob: -12.45

Results: Failure Cases

Not possible to understand why the method failed



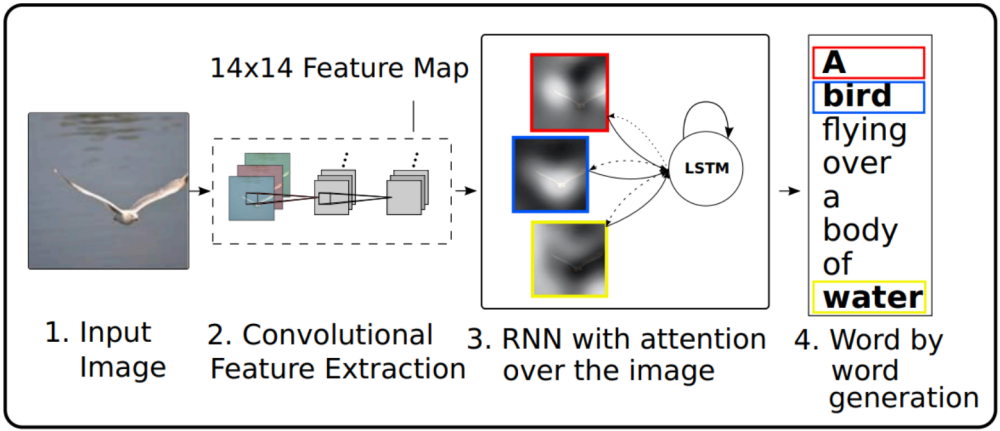
a woman holding a teddy bear in front of a mirror
logprob: -9.65

a horse is standing in the middle of a road
logprob: -10.34

How can we mitigate these failures?

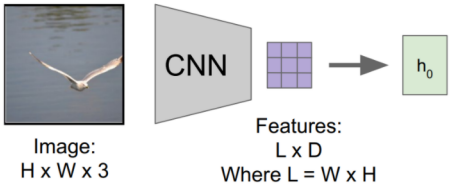
Image captioning **with attention**

Image Captioning with Attention



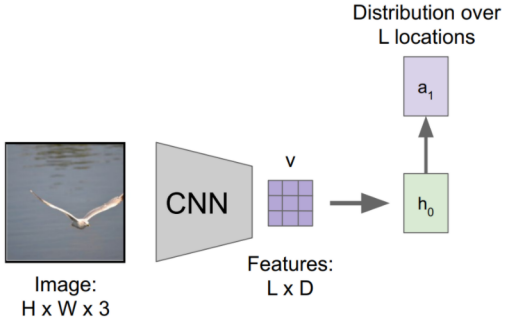
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



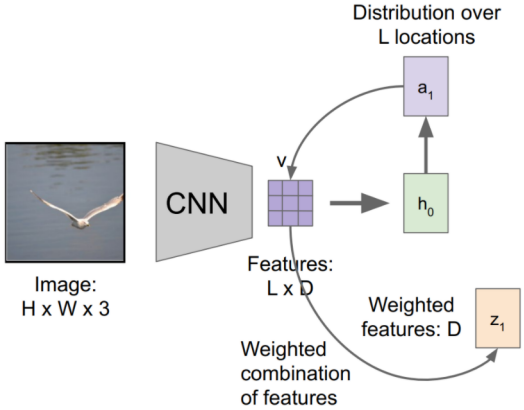
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

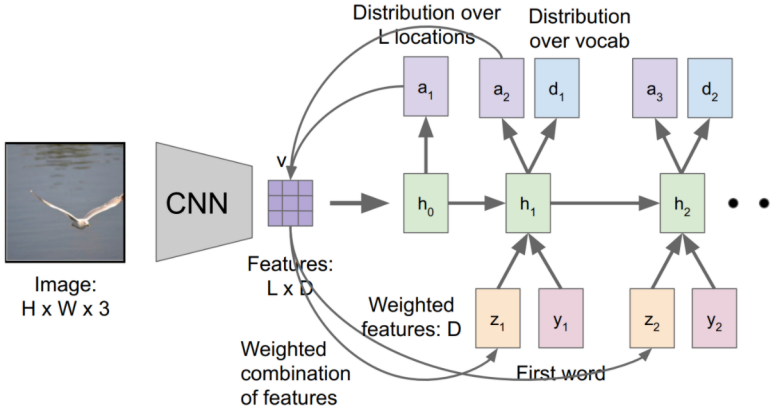
Image Captioning with Attention



$$z_1 = \sum_{i=1}^{L} a_i v_i$$

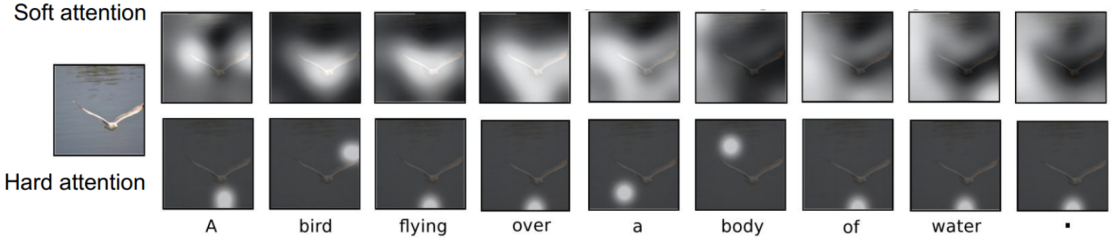
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention: Results

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

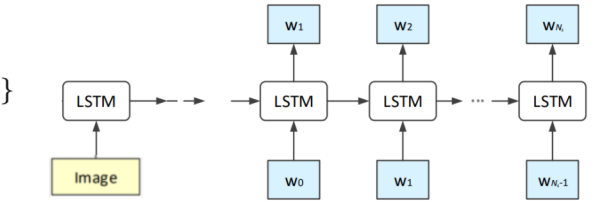
A giraffe standing in a forest with trees in the background.

Recent Efforts: Boosting Image Captioning with Attributes in LSTMs

$$\mathbf{x}^{-1} = \mathbf{T}_v \mathbf{I}\,, \quad \mathbf{x}^t = \mathbf{T}_s \mathbf{w}_t,$$

$$\mathbf{h}^t = f\left(\mathbf{x}^t\right),$$

$$t \in \{0, \ldots, N_s - 1\}$$



$\mathbf{I} \in \mathbb{R}^{D_v}$

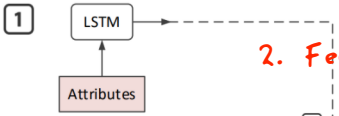
an image of a man in a boat with a dog

$\mathbf{w}_t \in \mathbb{R}^{D_s}$

$\{w_1, w_2, ..., w_{N_s}\}$

Boosting Image Captioning with Attributes in LSTMs: A1

1. Transforming the images into attributes

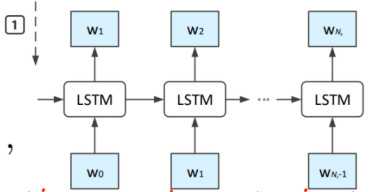


2. Feeding those attributes as inputs

$$\mathbf{x}^{-1} = \mathbf{T}_a \mathbf{A},$$

$$\mathbf{x}^{t} = \mathbf{T}_s \mathbf{w}_t,$$

$$\mathbf{h}^{t} = f\left(\mathbf{x}^{t}\right), \quad t \in \{0, \ldots, N_s - 1\},$$



why? bcoz the output gonna be a caption so why not giv textual data

$$\mathbf{A} \in \mathbb{R}^{D_a}$$

$$\mathcal{A} = \{a_1, a_2, ..., a_{D_a}\}$$

Attributes:
boat: 1 water: 0.838 man: 0.762
riding: 0.728 dog: 0.547 small: 0.485
person: 0.471 river: 0.461

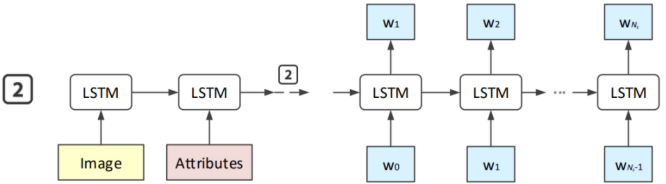
Credit: Yao et al, Boosting Image Captioning with Attributes, ICCV 2017

Boosting Image Captioning with Attributes in LSTMs: A2

$$\mathbf{x}^{-2} = \mathbf{T}_v \mathbf{I} \text{ and } \mathbf{x}^{-1} = \mathbf{T}_a \mathbf{A},$$

First: Image input
Second: Attribute input

$$\mathbf{x}^t = \mathbf{T}_s \mathbf{w}_t,$$



$$\mathbf{h}^t = f\left(\mathbf{x}^t\right), \quad t \in \{0, \ldots, N_s - 1\}$$

Credit: Yao et al, Boosting Image Captioning with Attributes, ICCV 2017

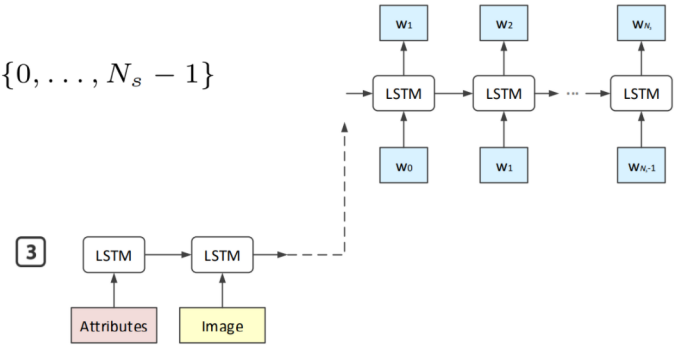
Boosting Image Captioning with Attributes in LSTMs: A3

$$\mathbf{x}^{-2} = \mathbf{T}_a \mathbf{A} \quad \text{and} \quad \mathbf{x}^{-1} = \mathbf{T}_v \mathbf{I},$$

$$\mathbf{x}^t = \mathbf{T}_s \mathbf{w}_t,$$

$$\mathbf{h}^t = f\left(\mathbf{x}^t\right), \quad t \in \{0, \ldots, N_s - 1\}$$

First: Attribute
Second: Image



Credit: Yao et al, Boosting Image Captioning with Attributes, ICCV 2017

Boosting Image Captioning with Attributes in LSTMs: A4

Attributes are given first only once

each time step, both image and words in caption are given as input

word

captions 4

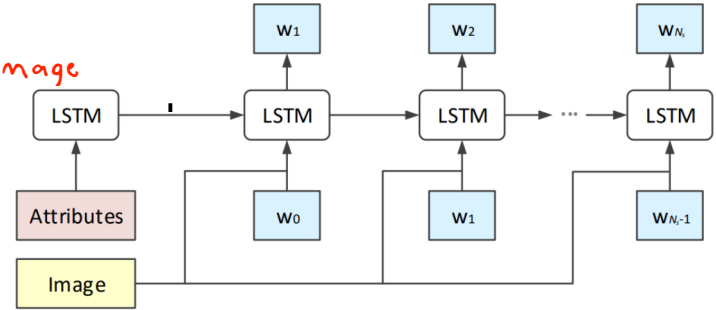
Image

$$\mathbf{x}^{-1} = \mathbf{T}_a \mathbf{A},$$

$$\mathbf{x}^t = \mathbf{T}_s \mathbf{w}_t + \mathbf{T}_v \mathbf{I},$$

$$\mathbf{h}^t = f\left(\mathbf{x}^t\right),$$

$$t \in \{0, \ldots, N_s - 1\}$$



Credit: Yao et al, Boosting Image Captioning with Attributes, ICCV 2017

Boosting Image Captioning with Attributes in LSTMs: A5

opposite of A4

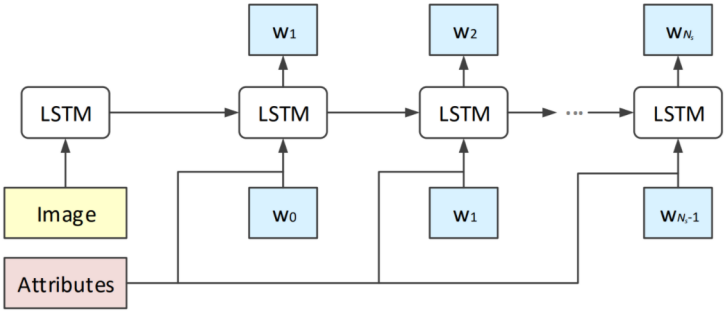
5

$$\mathbf{x}^{-1} = \mathbf{T}_v \mathbf{I},$$

$$\mathbf{x}^{t} = \mathbf{T}_s \mathbf{w}_t + \mathbf{T}_a \mathbf{A},$$

$$\mathbf{h}^{t} = f\left(\mathbf{x}^{t}\right),$$

$$t \in \{0, \ldots, N_s - 1\}$$

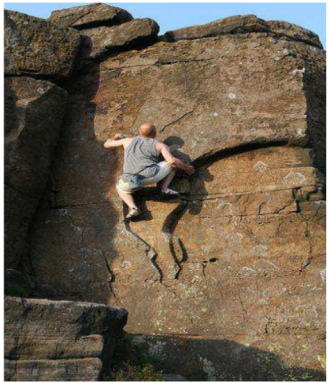


Credit: Yao et al, Boosting Image Captioning with Attributes, ICCV 2017

Boosting Image Captioning with Attributes in LSTMs: Observations

- LSTM-A_1 > LSTM
  - Indicates advantage of exploiting high-level attributes than image representations
- LSTM-A_2 > LSTM-A_1
  - Integrating image representations performs better
- LSTM-A_3 > LSTM-A_2
  - Benefits from mechanism of first feeding high-level attributes into LSTM instead of starting from image representations
- LSTM-A_4 < LSTM-A_3
  - This may be because noise in image can be explicitly accumulated, and thus network overfits more easily
  - But LSTM-A_5 which feeds attributes at each time step shows improvements on LSTM-A_3
- LSTM-A_2, LSTM-A_3, LSTM-A_5 > LSTM
  - Indicates that image representations and attributes are complementary and have mutual reinforcement for image captioning

StyleNet: Generating Attractive Visual Captions with Styles



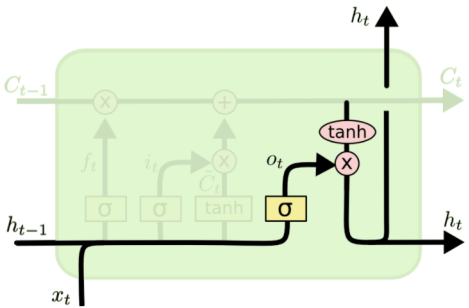
CaptionBot: A man on a rocky hillside next to a stone wall.

Romantic: A man uses rock climbing to conquer the high.

Humorous: A man is climbing the rock like a lizard.

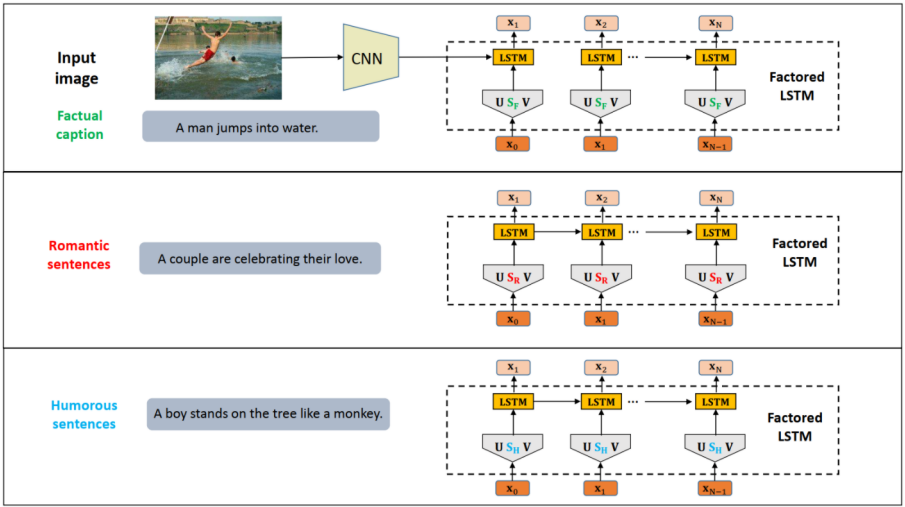
StyleNet: Generating Attractive Visual Captions with Styles

StyleNet proposes a factored LSTM

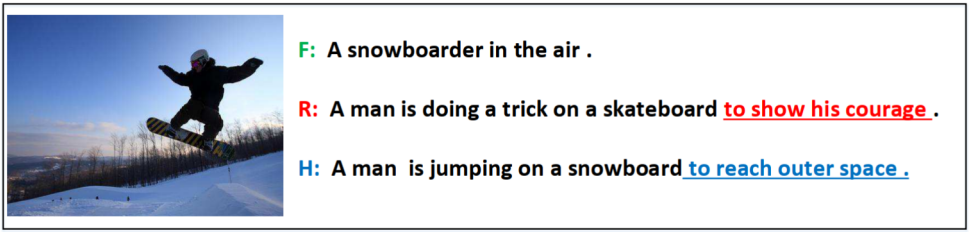


$$\boldsymbol{W}_x = \boldsymbol{U}_x \boldsymbol{S}_x \boldsymbol{V}_x$$
$$\boldsymbol{i}_t = \text{sigmoid}(\boldsymbol{W}_{ix}\boldsymbol{x}_t + \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1})$$
$$\boldsymbol{f}_t = \text{sigmoid}(\boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1})$$
$$\boldsymbol{o}_t = \text{sigmoid}(\boldsymbol{W}_{ox}\boldsymbol{x}_t + \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1})$$
$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_{cx}\boldsymbol{x}_t + \boldsymbol{W}_{ch}\boldsymbol{h}_{t-1})$$
$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t$$
$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \boldsymbol{c}_t$$

StyleNet: Generating Attractive Visual Captions with Styles

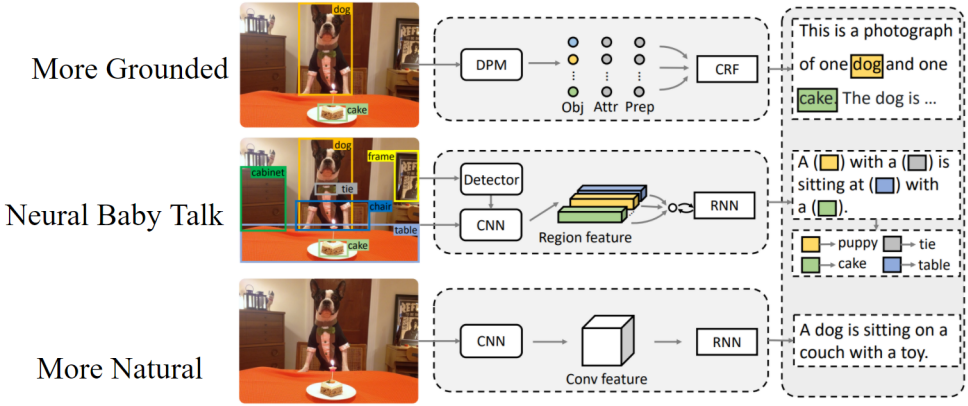


StyleNet: Generating Attractive Visual Captions with Styles



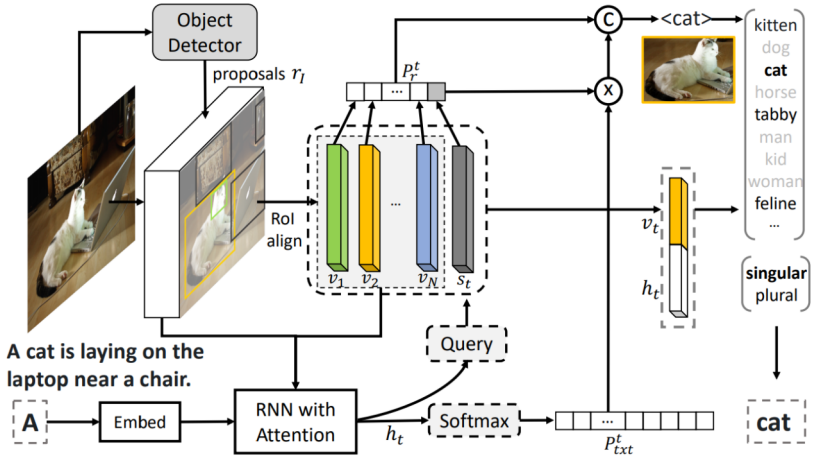
At test swap S_x accordingly to get the desired output type.

Neural Baby Talk



Credit: Lu et al, Neural Baby Talk, CVPR 2018

Neural Baby Talk



Credit: Lu et al, Neural Baby Talk, CVPR 2018