

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!



Review: Are autoencoders (AE) and PCA connected?

- Yes! Consider normalizing the inputs to the AE as:

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left(x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$



NPTEL

Review: Are autoencoders (AE) and PCA connected?

- Yes! Consider normalizing the inputs to the AE as:

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left(x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$



NPTEL

- What is this doing? Making the mean of each dimension zero

Review: Are autoencoders (AE) and PCA connected?

- Yes! Consider normalizing the inputs to the AE as:

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left(x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$



NPTEL

- What is this doing? Making the mean of each dimension zero

- Then, covariance matrix is:

$$\frac{1}{m} \hat{X}^T \hat{X} = X^T X$$

Review: Are autoencoders (AE) and PCA connected?

- Yes! Consider normalizing the inputs to the AE as:

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left(x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$

- What is this doing? Making the mean of each dimension zero
- Then, covariance matrix is:

$$\frac{1}{m} \hat{X}^T \hat{X} = X^T X$$

- Now, consider AE loss:

$$\min_{\theta} \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \hat{x}_{ij})^2 \longleftrightarrow \min_{W^* H} (||X - HW^*||_F)^2$$

where $||A||_F = \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2$ is Frobenius norm, H is AE representation and W^* are decoder weights

The NPTEL logo features the letters "NPTEL" in a bold, sans-serif font. The letters are colored in a gradient from light orange to yellow. Behind the letters is a stylized circular emblem consisting of concentric rings in shades of brown and tan, with a central flower-like shape.

Review: Are autoencoders (AE) and PCA connected?

- Yes! Consider normalizing the inputs to the AE as:

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left(x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$

- What is this doing? Making the mean of each dimension zero
- Then, covariance matrix is:
 $\frac{1}{m} \hat{X}^T \hat{X} = X^T X$

- Now, consider AE loss:

$$\min_{\theta} \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \hat{x}_{ij})^2 \longleftrightarrow \min_{W^* H} (\|X - HW^*\|_F)^2$$

where $\|A\|_F = \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2$ is Frobenius norm, H is AE representation and W^* are decoder weights

- From SVD, we know the optimal solution is:
 $HW^* = U_{\cdot, \leq k} \Sigma_{k,k} V_{\cdot, \leq k}^T$

Review: Are autoencoders (AE) and PCA connected?

- Yes! Consider normalizing the inputs to the AE as:

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left(x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$

- What is this doing? Making the mean of each dimension zero
- Then, covariance matrix is:

$$\frac{1}{m} \hat{X}^T \hat{X} = X^T X$$

- Now, consider AE loss:

$$\min_{\theta} \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \hat{x}_{ij})^2 \longleftrightarrow \min_{W^* H} (\|X - HW^*\|_F)^2$$

where $\|A\|_F = \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2$ is Frobenius norm, H is AE representation and W^* are decoder weights

- From SVD, we know the optimal solution is:
$$HW^* = U_{\cdot, \leq k} \Sigma_{k,k} V_{\cdot, \leq k}^T$$
- One possible solution then is:

$$H = U_{\cdot, \leq k} \Sigma_{k,k}$$
$$W^* = V_{\cdot, \leq k}^T$$

Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$H = U_{\cdot, \leq k} \Sigma_{k,k}$$



Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T) (X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \end{aligned}$$

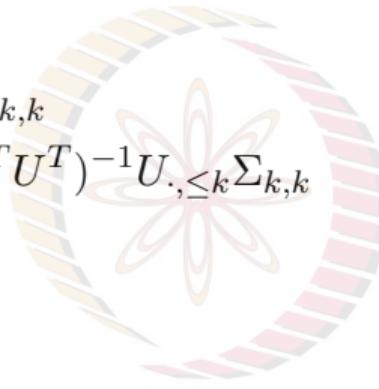


NPTEL

Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T) (X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T) (U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \end{aligned}$$

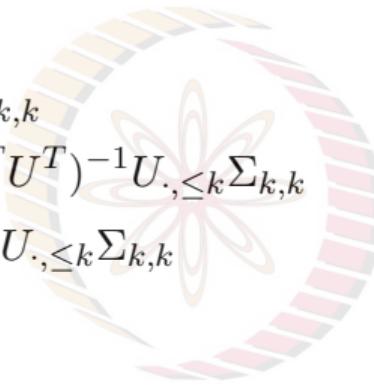


NPTEL

Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

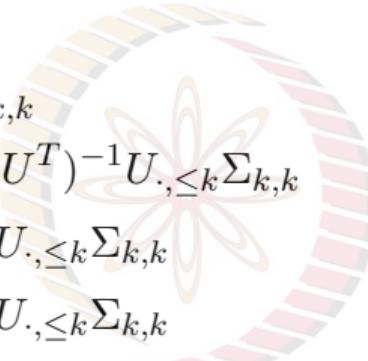
$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T) (X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T) (U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \end{aligned}$$



Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T) (X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T) (U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \end{aligned}$$

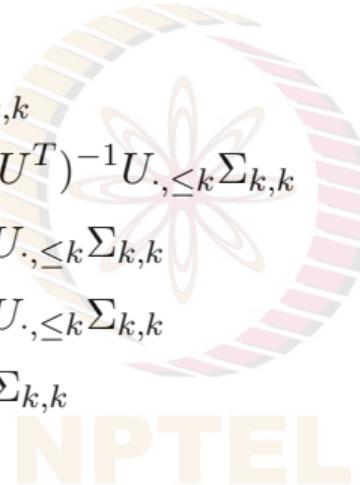


NPTEL

Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

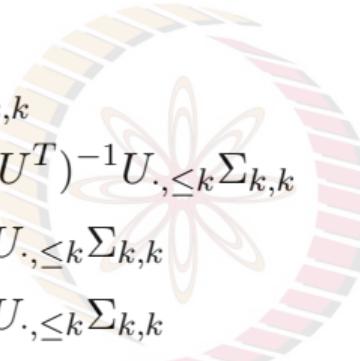
$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T) (X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T) (U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \end{aligned}$$



Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (XX^T)(XX^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (XV\Sigma^T U^T)(U\Sigma V^T V\Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= XV\Sigma^T U^T (U\Sigma\Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= XV\Sigma^T U^T U (\Sigma\Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= XV\Sigma^T (\Sigma\Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= XV\Sigma^T (\Sigma^T)^{-1} \Sigma^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \end{aligned}$$

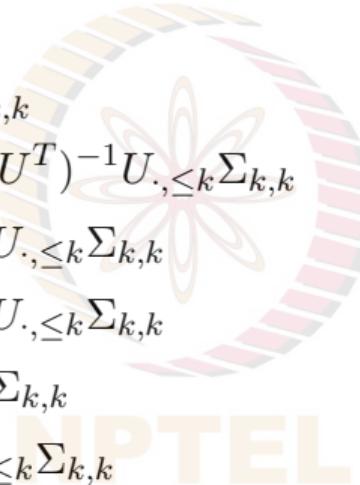


NPTEL

Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T)(X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T)(U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma^T)^{-1} \Sigma^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^{-1} I_{\cdot, \leq k} \Sigma_{k,k} \end{aligned}$$

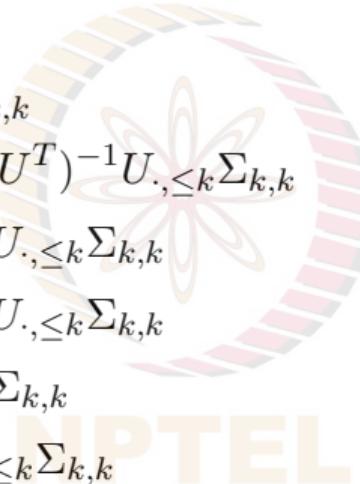


IIT-MADRAS

Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T)(X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T)(U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma^T)^{-1} \Sigma^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^{-1} I_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V I_{\cdot, \leq k} = X V_{\cdot, \leq k} \end{aligned}$$



Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T)(X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T)(U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma^T)^{-1} \Sigma^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^{-1} I_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V I_{\cdot, \leq k} = X V_{\cdot, \leq k} \end{aligned}$$

- H is a linear transformation of X and $W = V_{\cdot, \leq k}$!

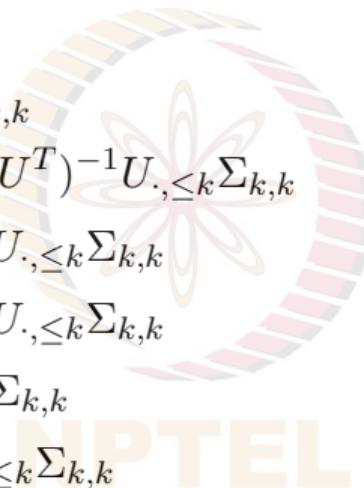
Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T)(X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T)(U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma^T)^{-1} \Sigma^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^{-1} I_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V I_{\cdot, \leq k} = X V_{\cdot, \leq k} \end{aligned}$$

- H is a linear transformation of X and $W = V_{\cdot, \leq k}$!

- From SVD, we know V is matrix of eigenvectors of $X^T X$, the covariance matrix



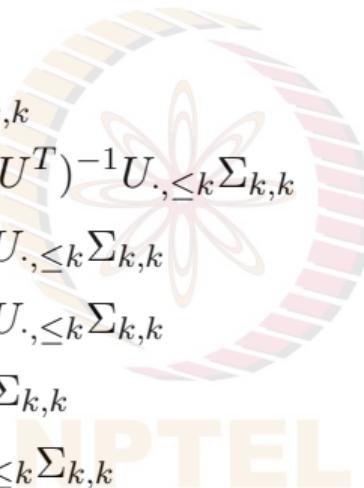
Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T)(X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T)(U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma^T)^{-1} \Sigma^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^{-1} I_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V I_{\cdot, \leq k} = X V_{\cdot, \leq k} \end{aligned}$$

- H is a linear transformation of X and $W = V_{\cdot, \leq k}$!

- From SVD, we know V is matrix of eigenvectors of $X^T X$, the covariance matrix
- The encoder weights are eigenvectors of covariance matrix, what does this tell you?



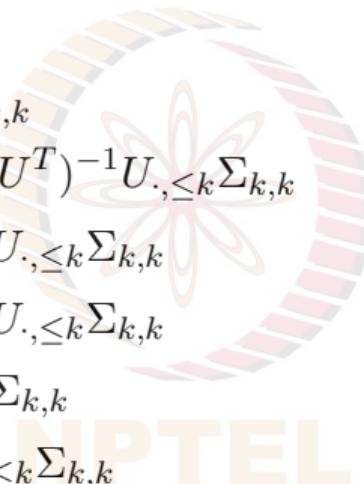
Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T)(X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T)(U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma^T)^{-1} \Sigma^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^{-1} I_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V I_{\cdot, \leq k} = X V_{\cdot, \leq k} \end{aligned}$$

- H is a linear transformation of X and $W = V_{\cdot, \leq k}$!

- From SVD, we know V is matrix of eigenvectors of $X^T X$, the covariance matrix
- The encoder weights are eigenvectors of covariance matrix, what does this tell you?
- This is PCA indeed! Is it so always?



Review: Are autoencoders (AE) and PCA connected?

- Then, we have:

$$\begin{aligned} H &= U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X X^T)(X X^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= (X V \Sigma^T U^T)(U \Sigma V^T V \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma \Sigma^T)^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^T (\Sigma^T)^{-1} \Sigma^{-1} U^T U_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V \Sigma^{-1} I_{\cdot, \leq k} \Sigma_{k,k} \\ &= X V I_{\cdot, \leq k} = X V_{\cdot, \leq k} \end{aligned}$$

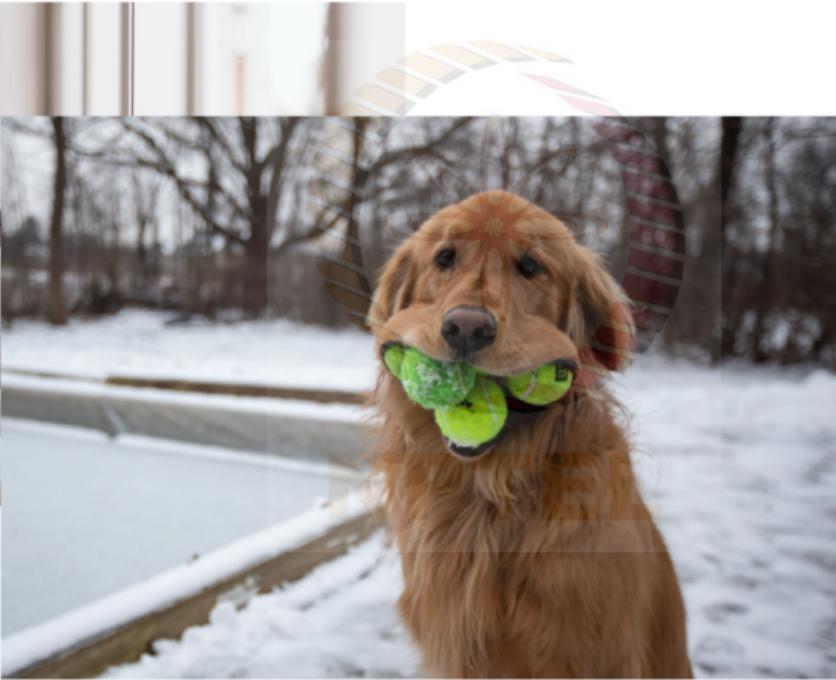
- H is a linear transformation of X and $W = V_{\cdot, \leq k}$!

- From SVD, we know V is matrix of eigenvectors of $X^T X$, the covariance matrix
- The encoder weights are eigenvectors of covariance matrix, what does this tell you?
- This is PCA indeed! Is it so always?
- No, when encoder and decoder are linear; inputs are normalized dimension-wise, as we saw; and we use MSE as loss function

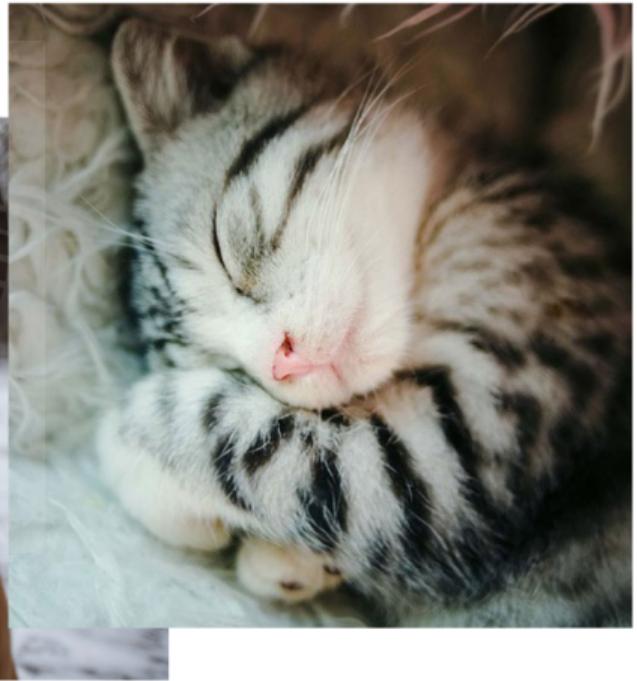
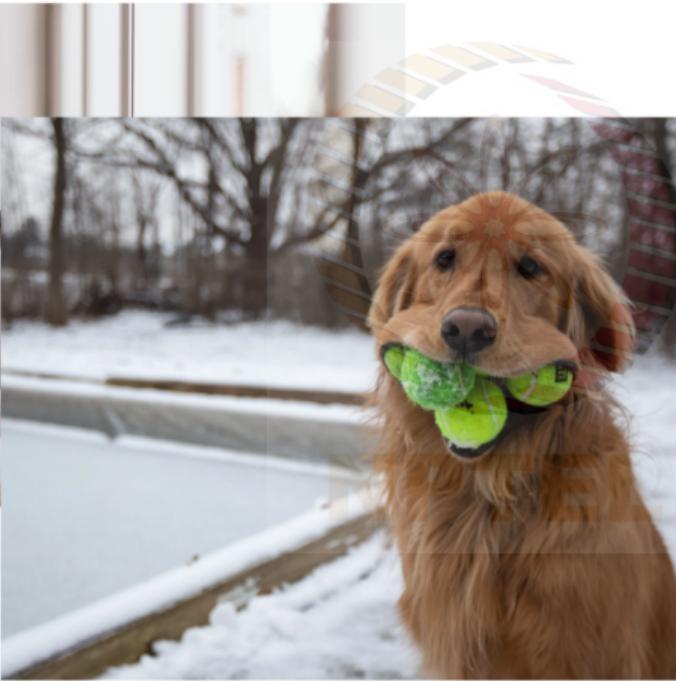
Describe these images



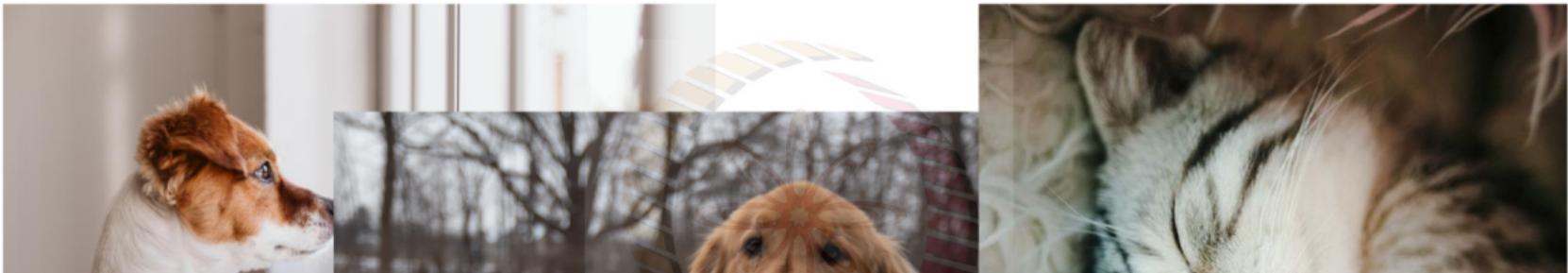
Describe these images



Describe these images



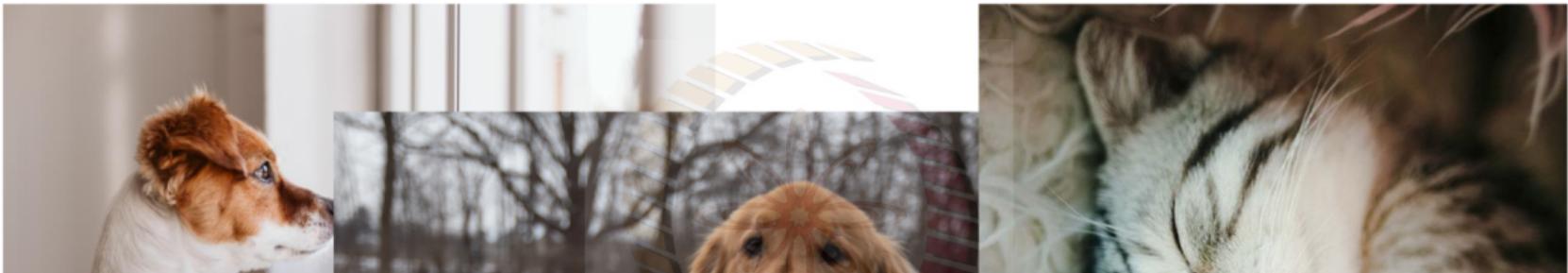
Describe these images



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



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?



Some Method

NPTEL

How to make a computer describe an image?

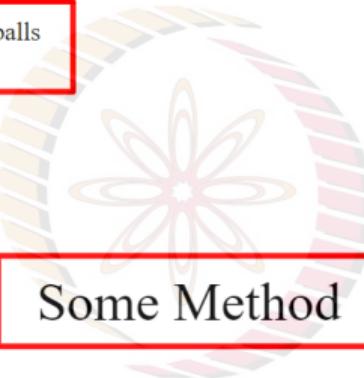


NPTEL

How to make a computer describe an image?



Dog is holding tennis balls
with its mouth.



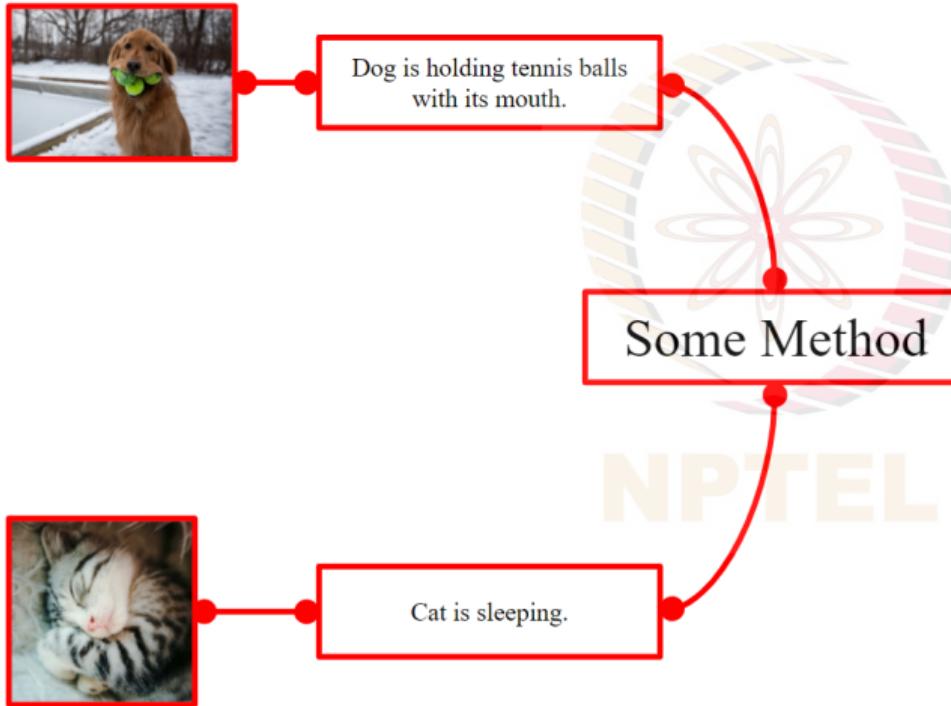
Some Method



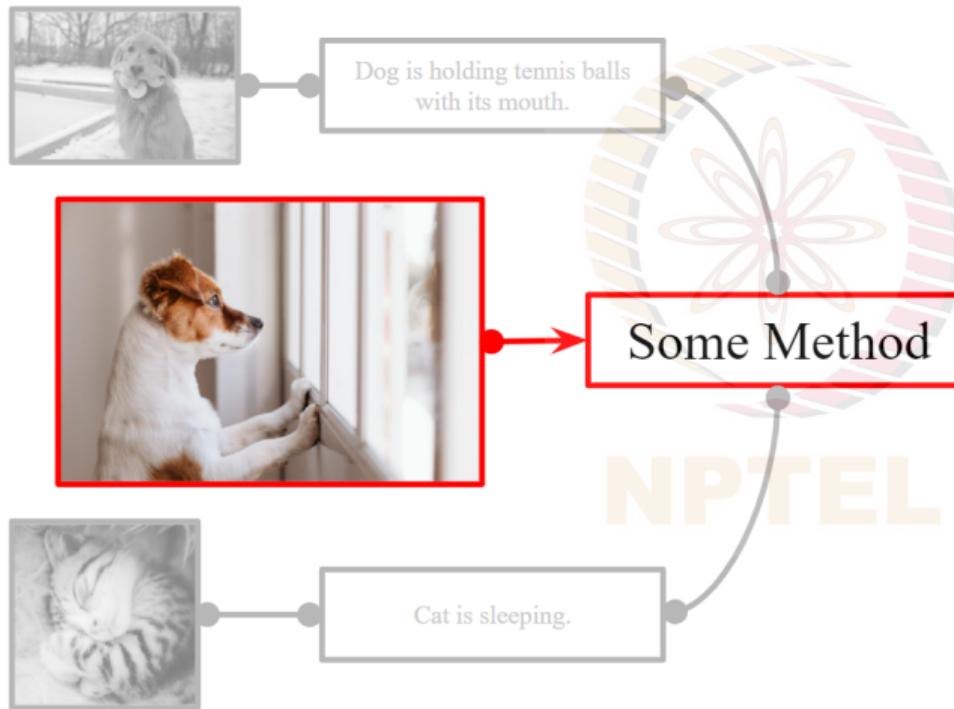
Cat is sleeping.

NPTEL

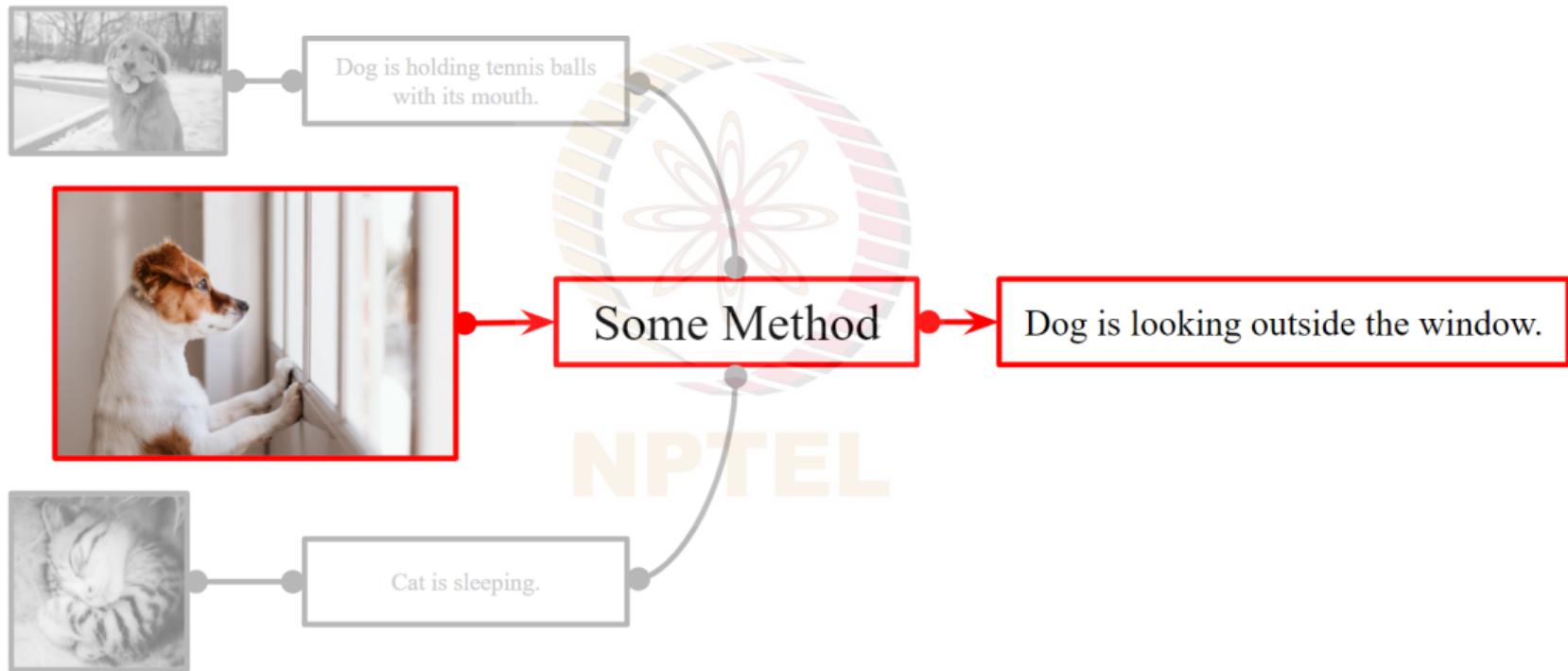
How to make a computer describe an image?



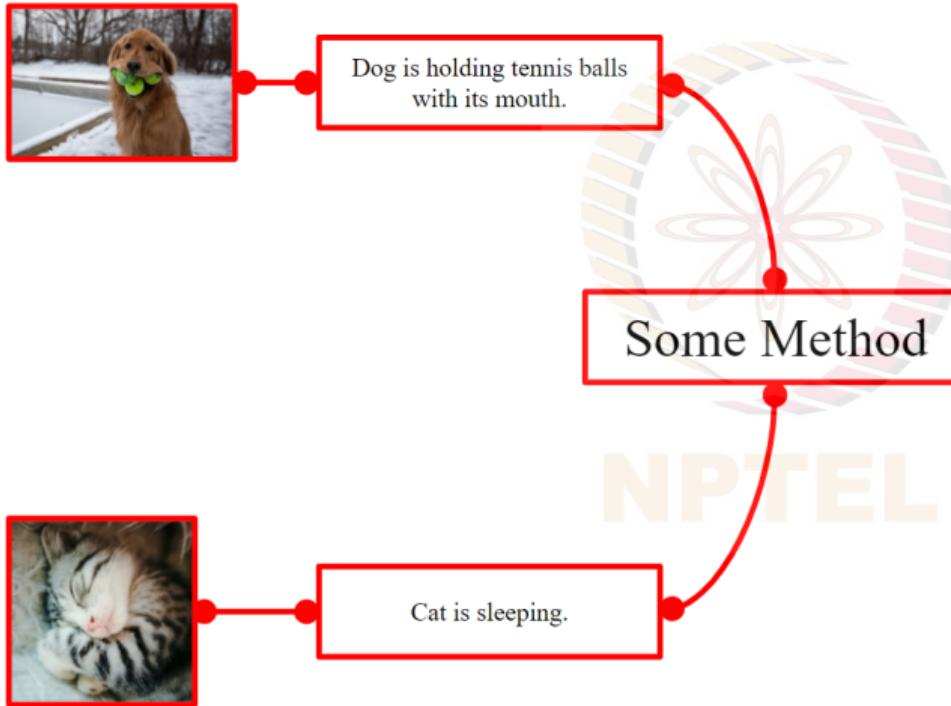
How to make a computer describe an image?



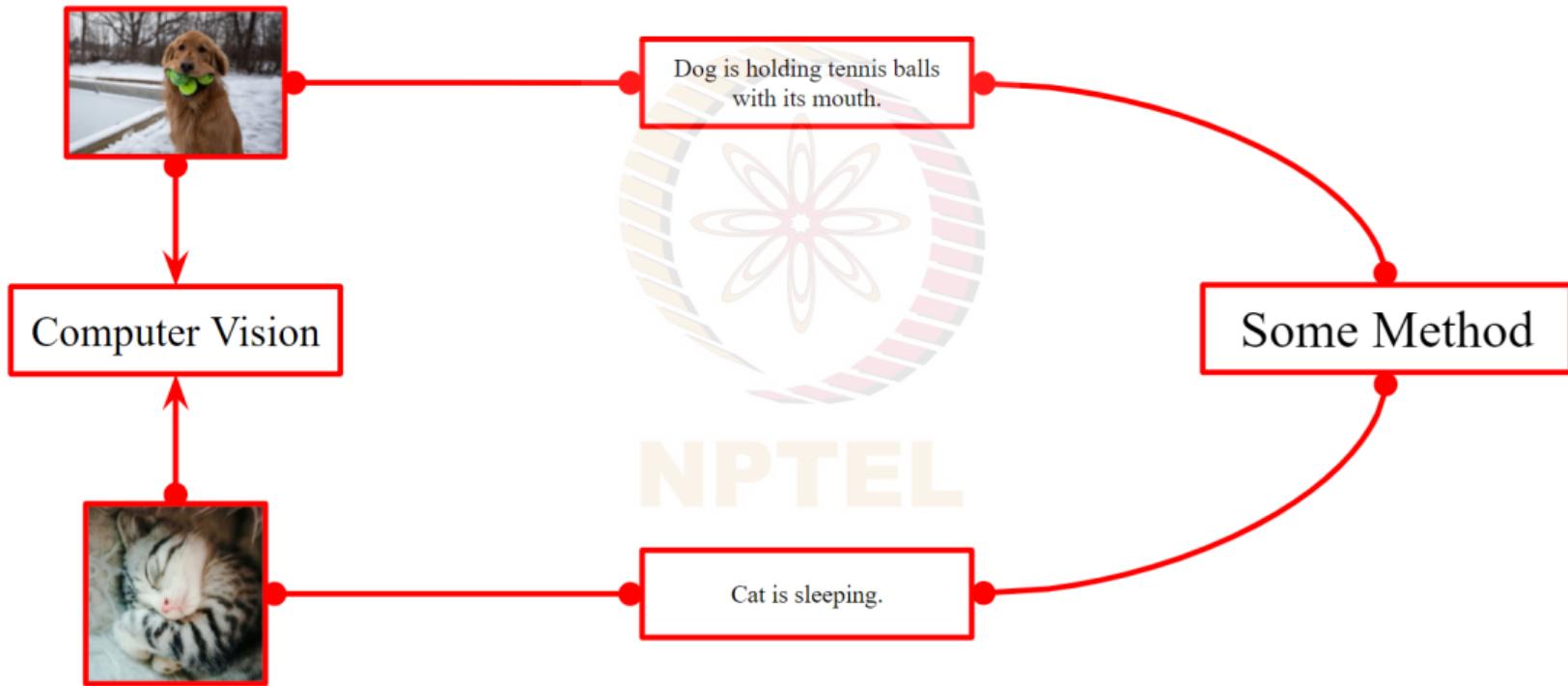
How to make a computer describe an image?



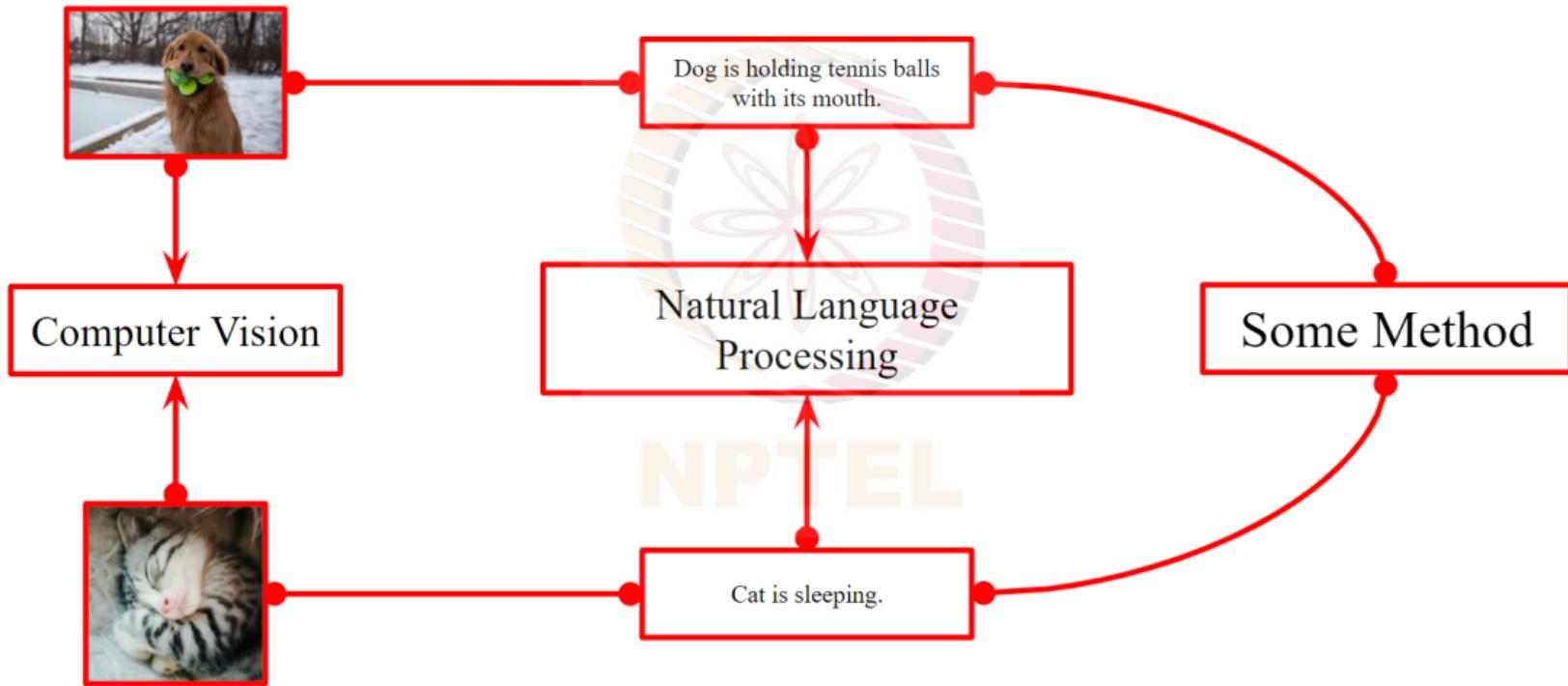
How to make a computer describe an image?



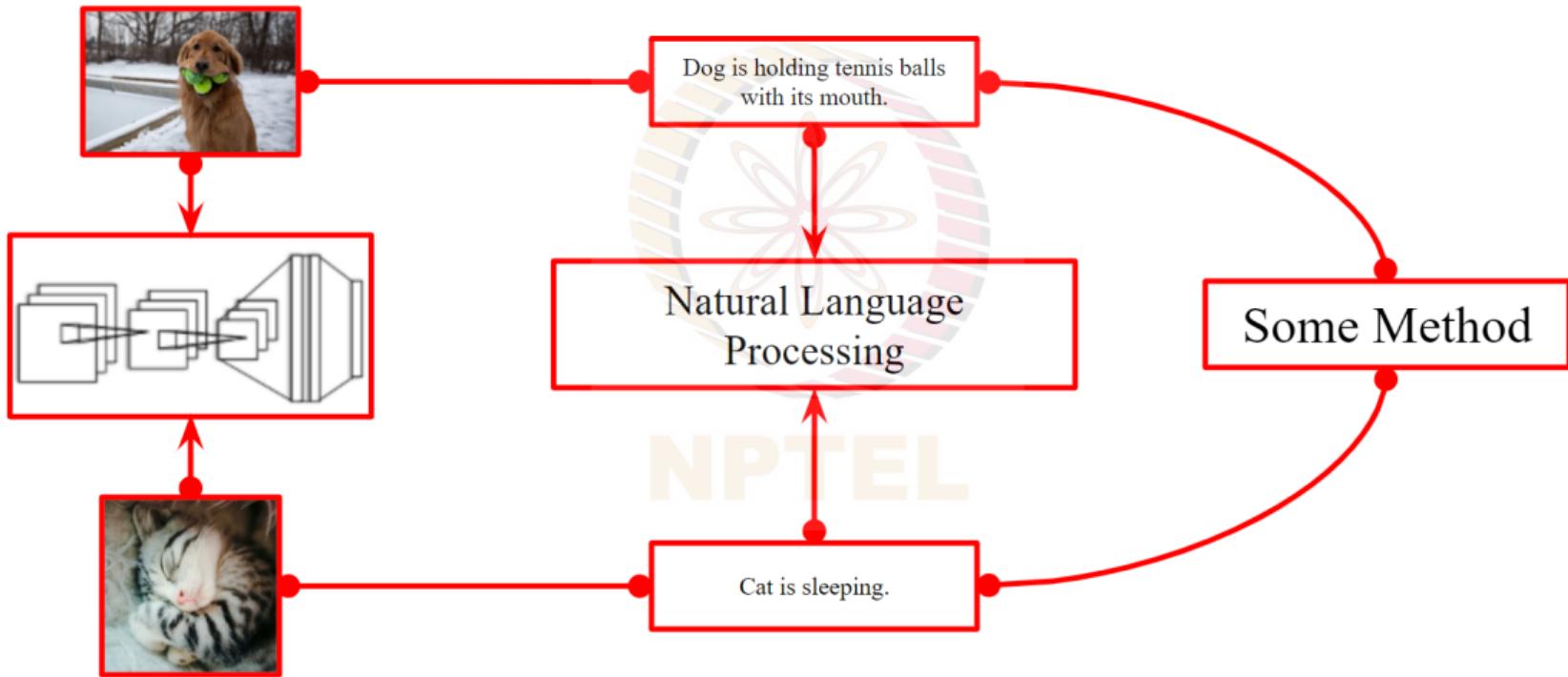
How to make a computer describe an image?



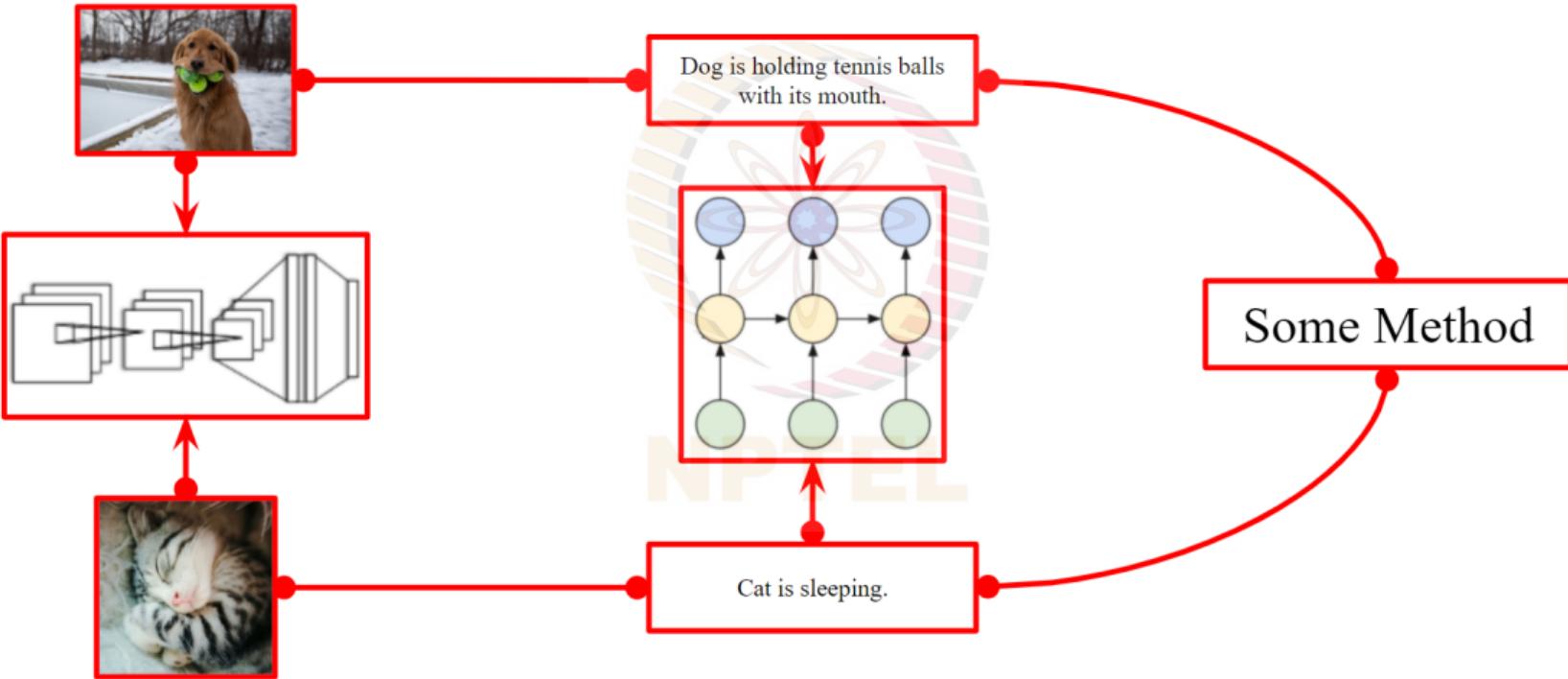
How to make a computer describe an image?



How to make a computer describe an image?



How to make a computer describe an image?



How to make a computer describe an image?

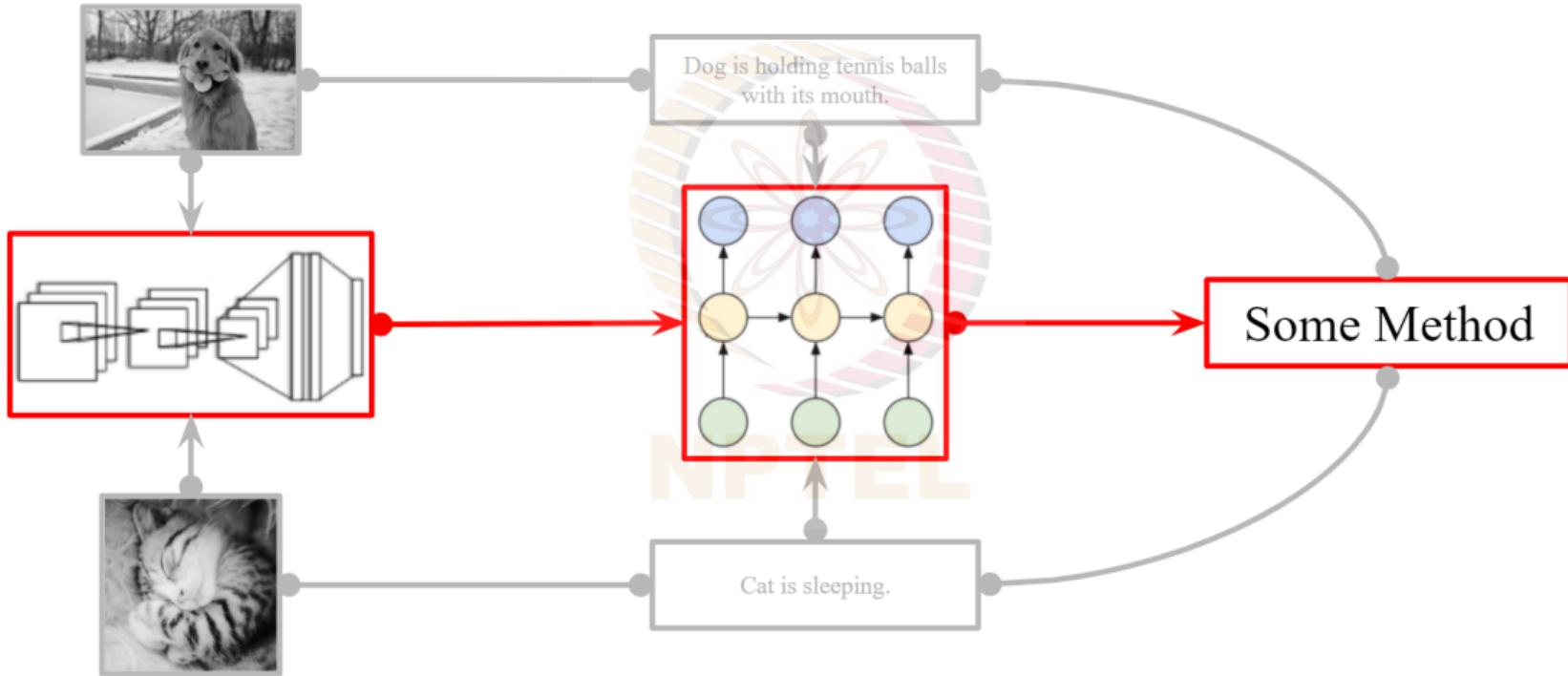


Image Captioning: Training



NPTEL



“straw hat”

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

Image Captioning: Training

image

conv-64

conv-64

maxpool

conv-128

conv-128

maxpool

conv-256

conv-256

maxpool

conv-512

conv-512

maxpool

conv-512

conv-512

maxpool

FC-4096

FC-4096

FC-1000

softmax



NPTEL

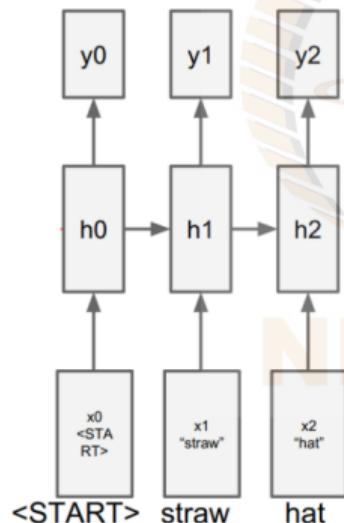


“straw hat”

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

Image Captioning: Training

image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax

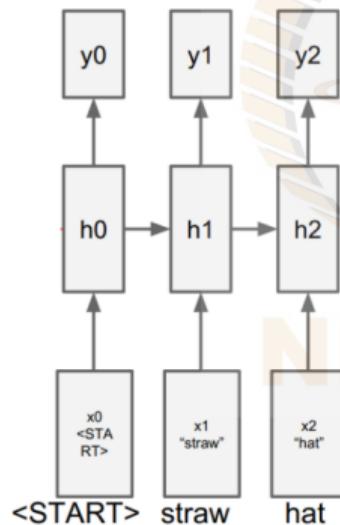


"straw hat"

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

Image Captioning: Training

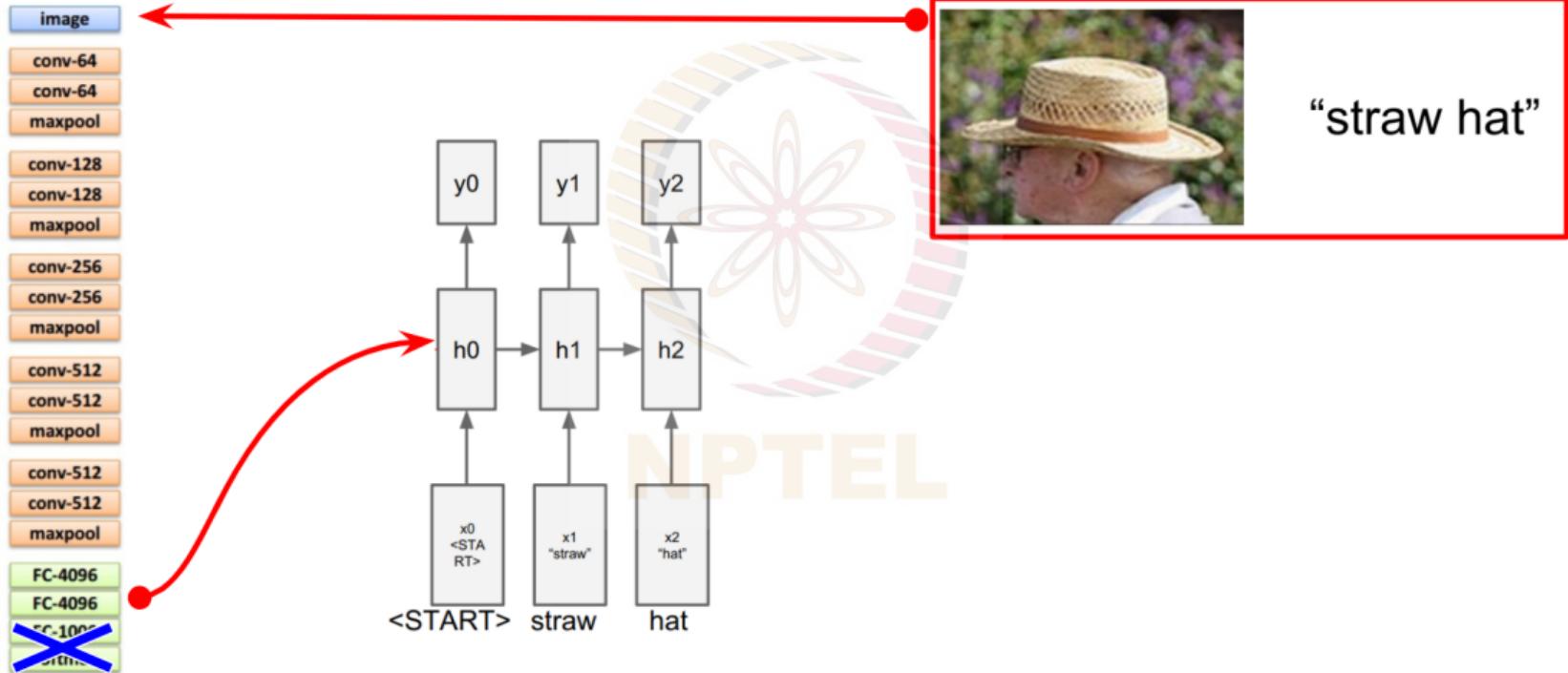
image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
~~FC-1000~~
~~softmax~~



"straw hat"

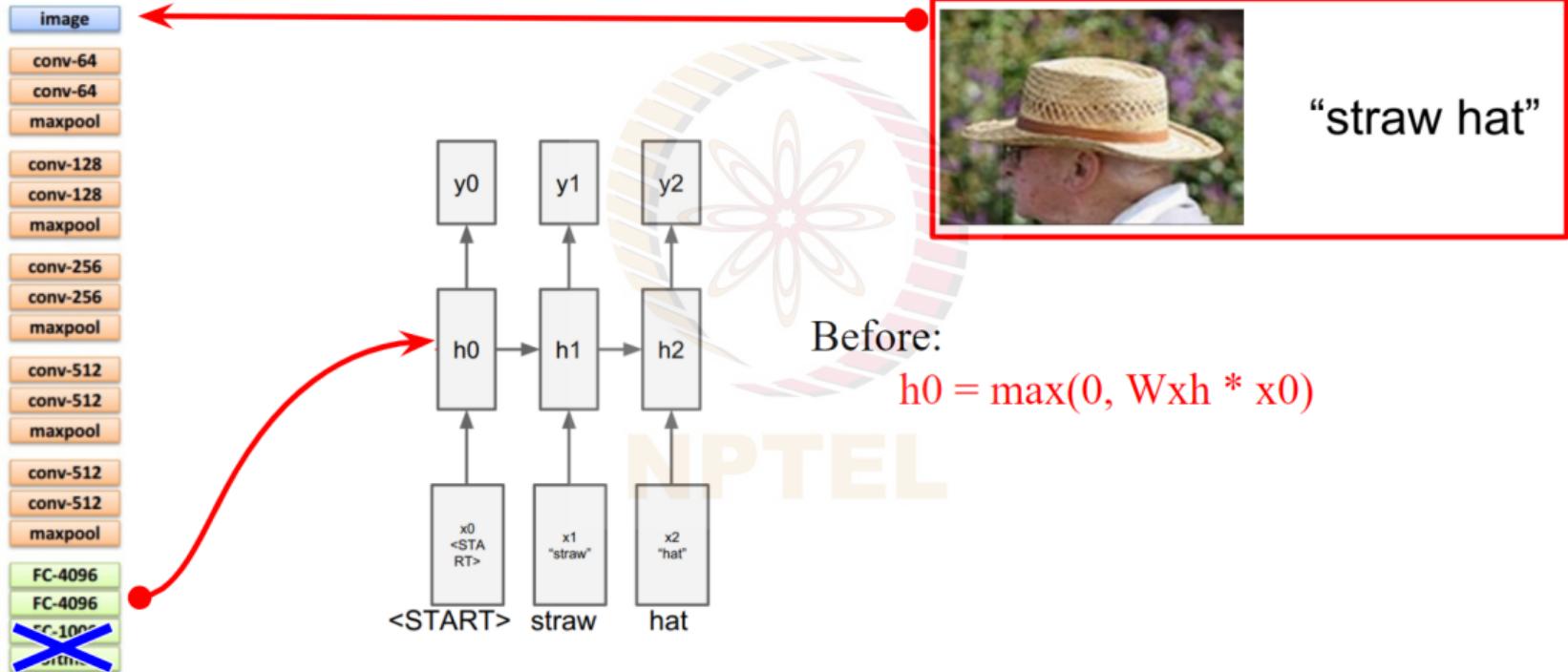
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Training



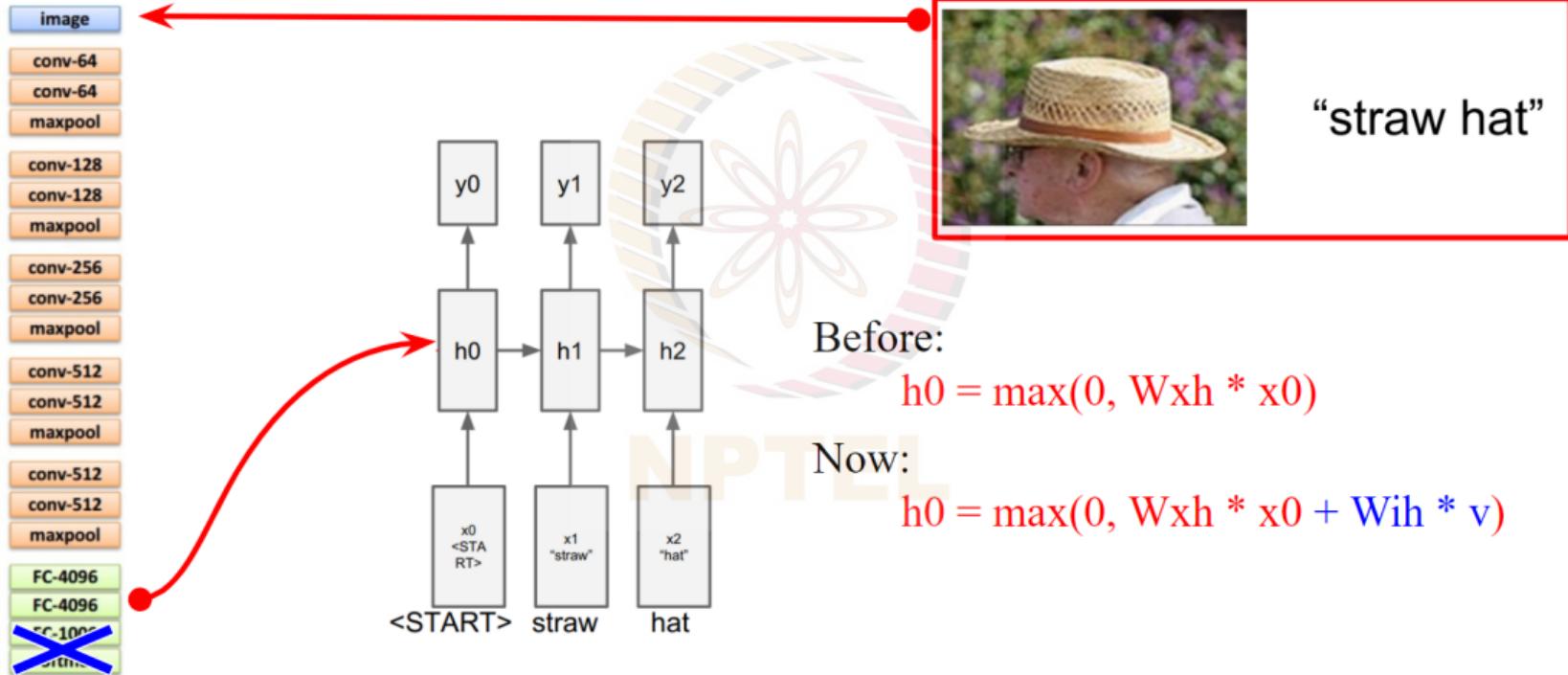
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Training



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

Image Captioning: Training



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

Image Captioning: Inference (Test Time)



NPTEL



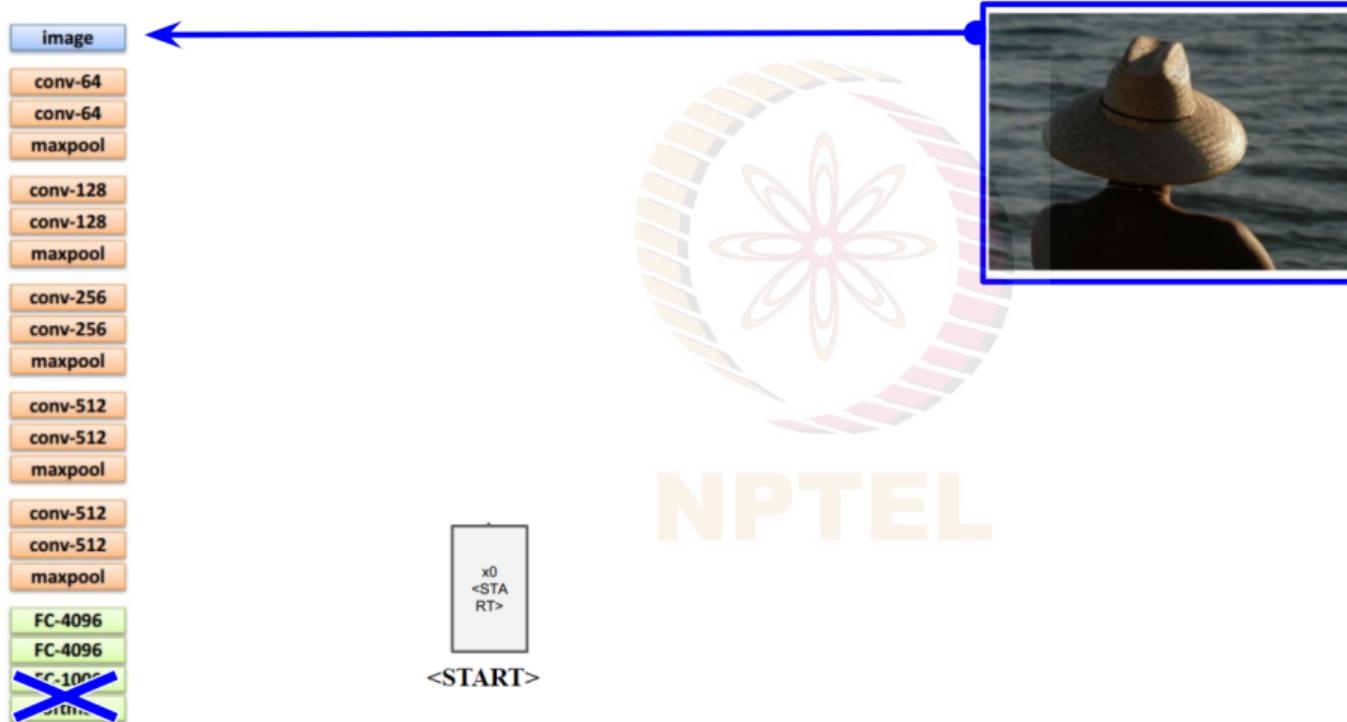
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



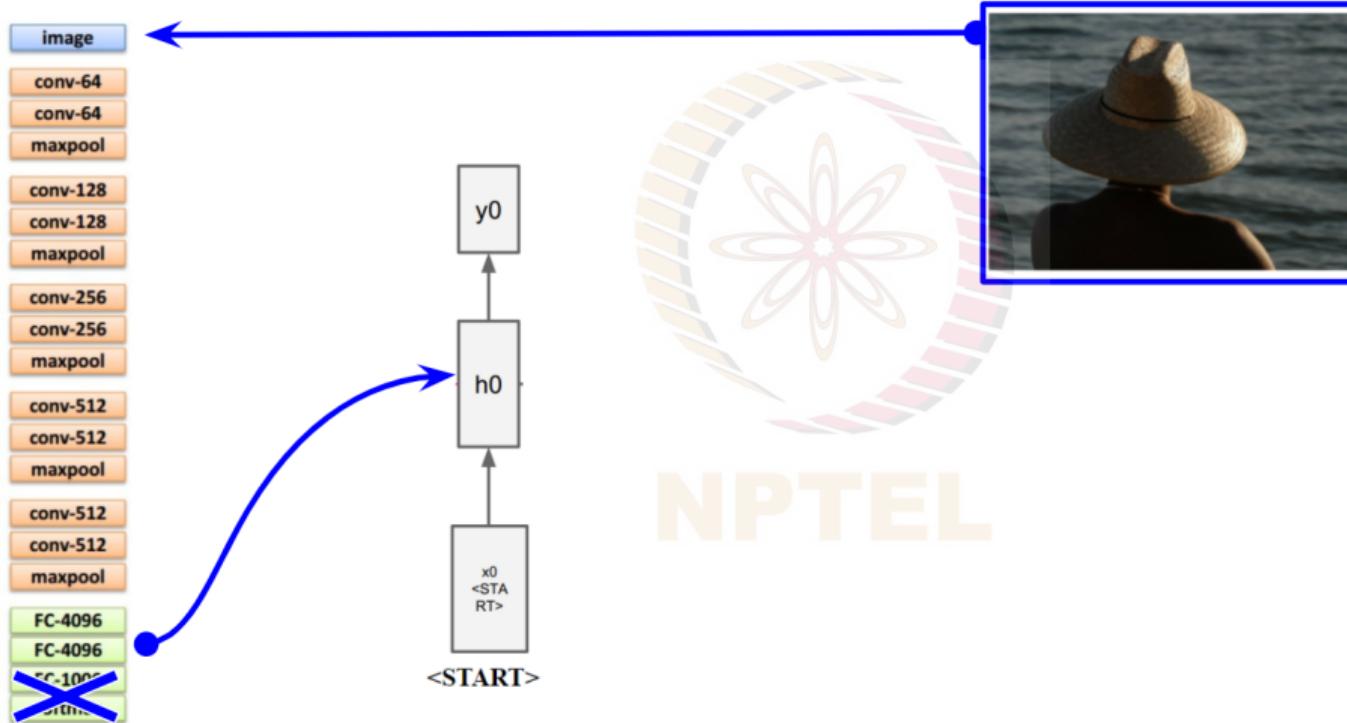
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



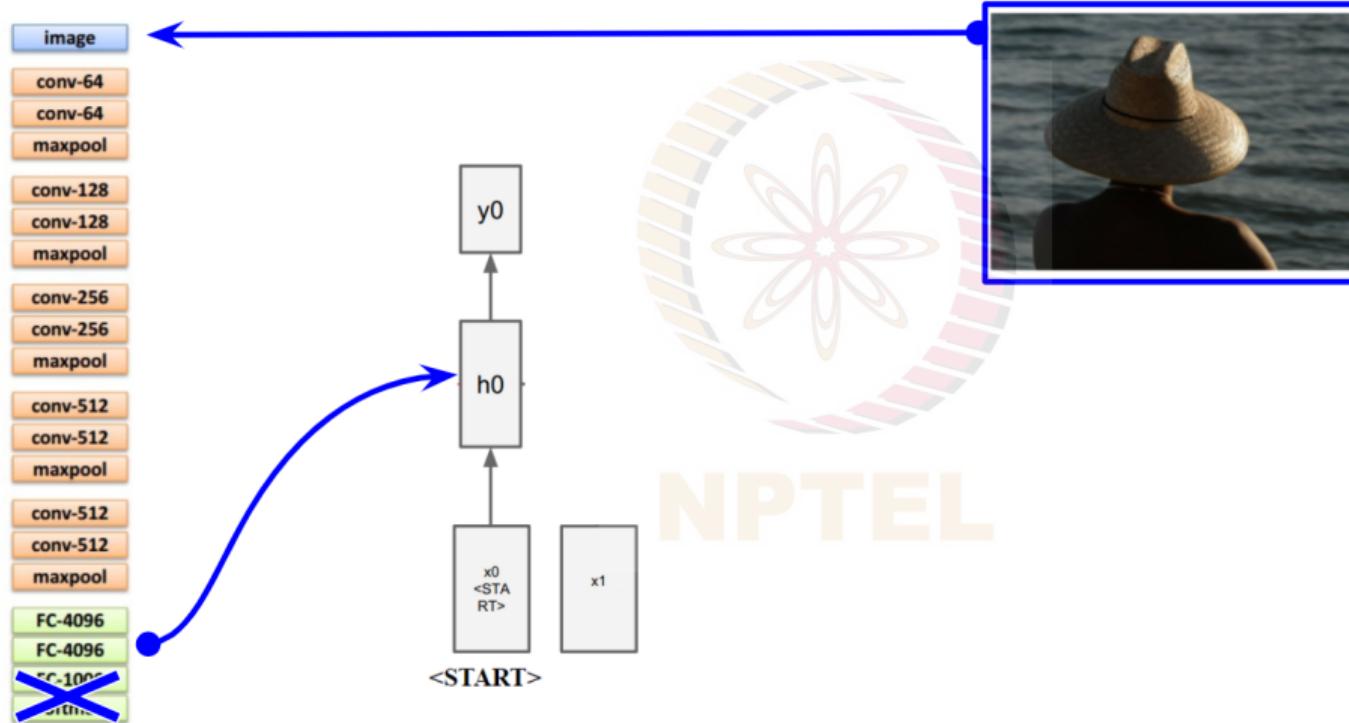
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



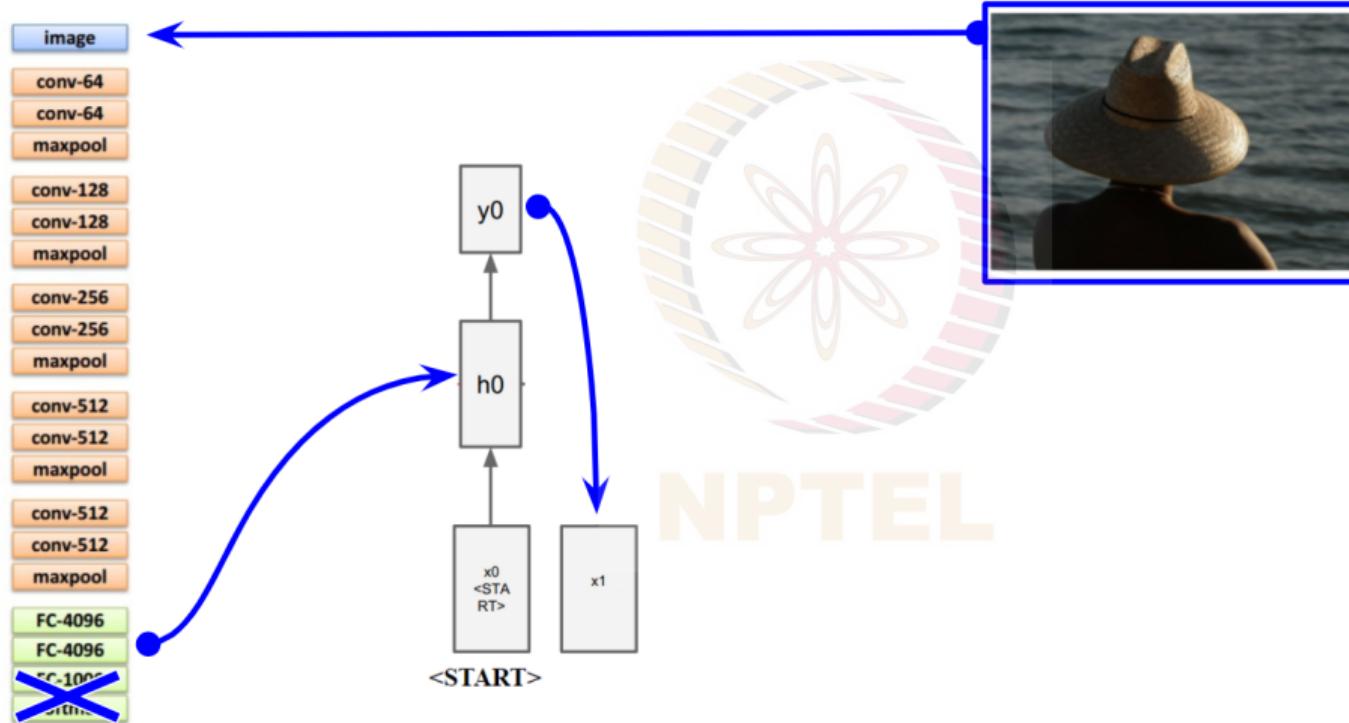
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



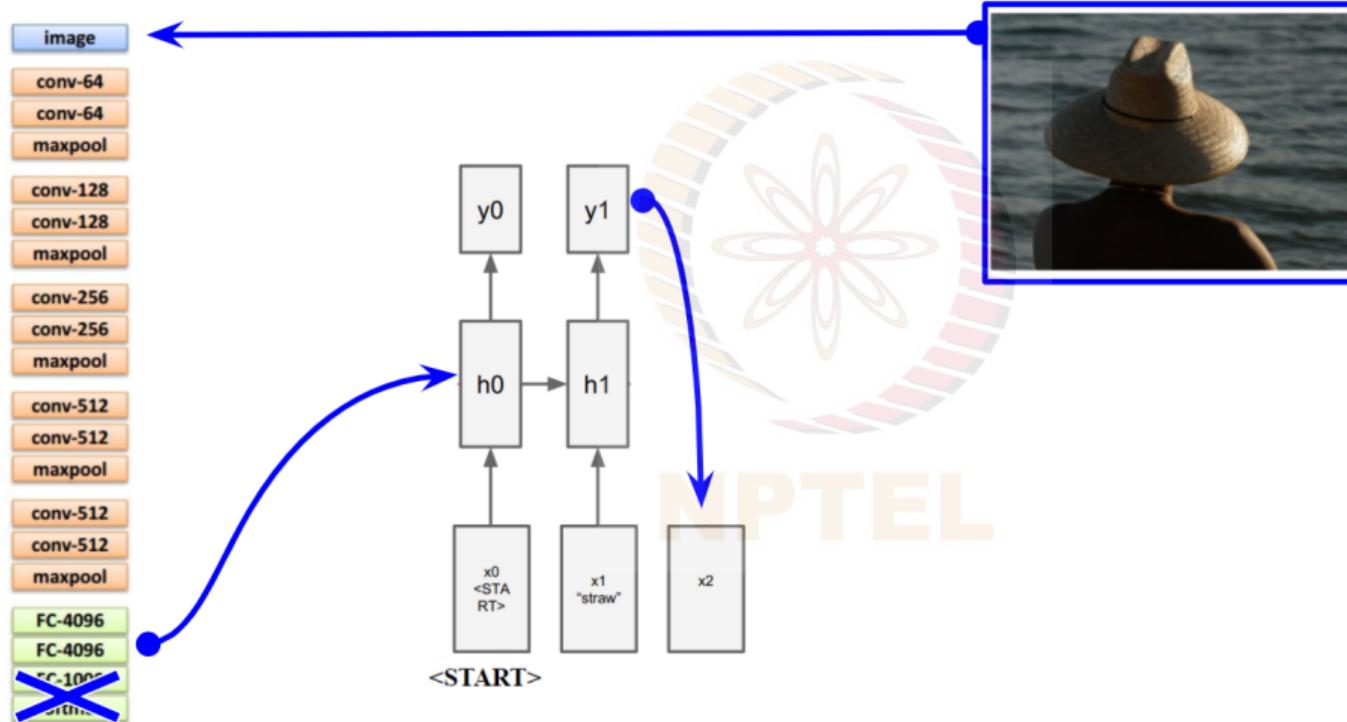
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



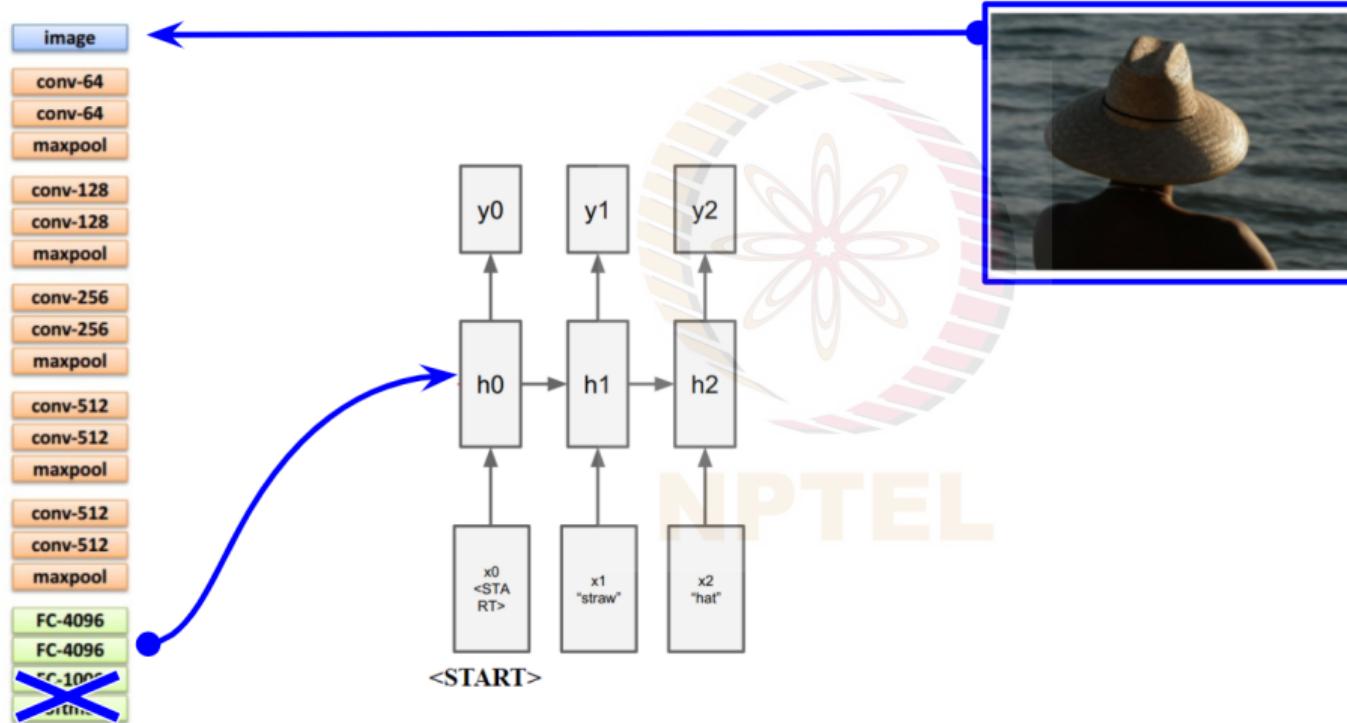
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



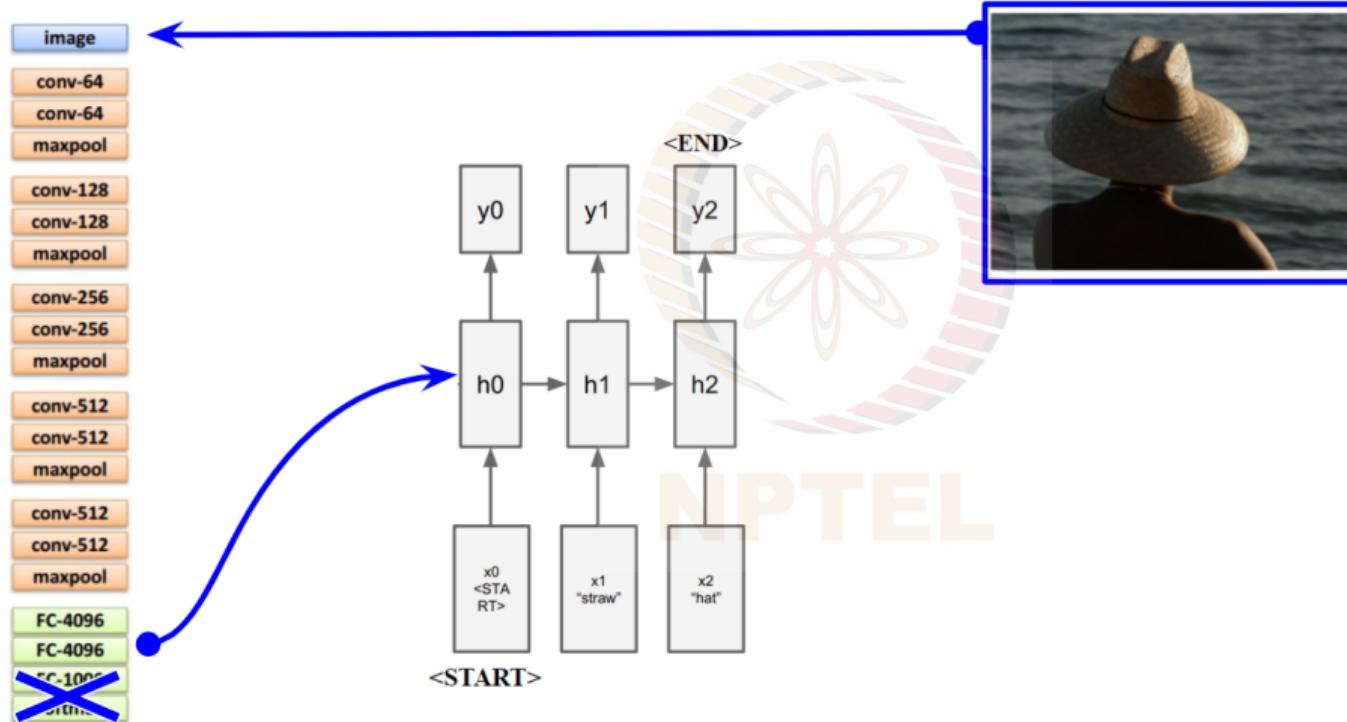
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

Image Captioning: Inference (Test Time)



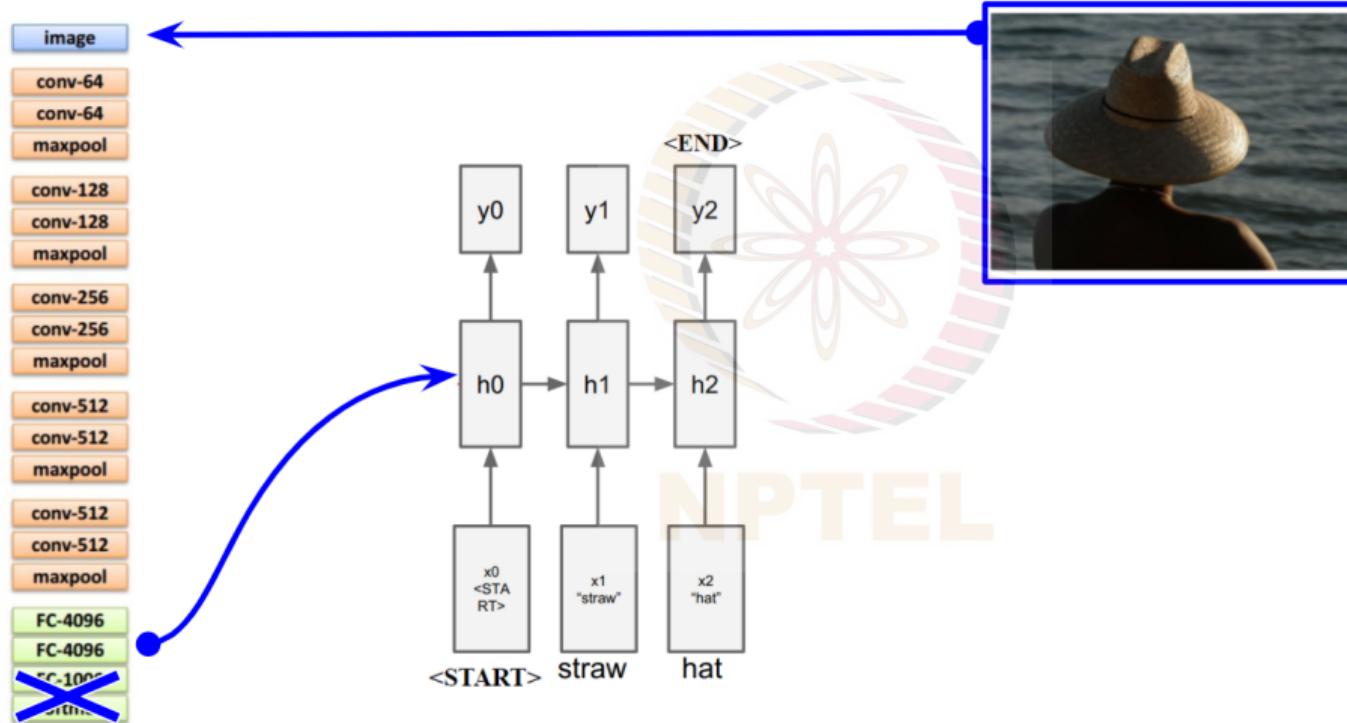
Credit: Karpathy et al, Deep visual-semantic alignments for generating image descriptions, CVPR 2015

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

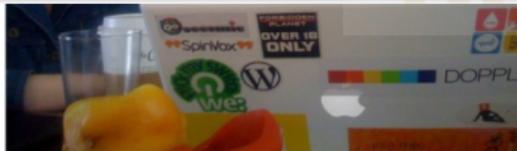


Results: Failure Cases

Not possible to understand why the method failed



How can we mitigate these failures?



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

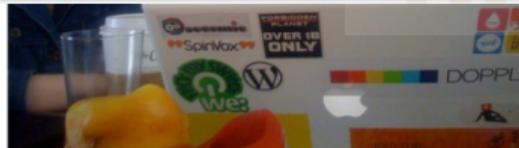
Results: Failure Cases

Not possible to understand why the method failed



How can we mitigate these failures?

Image captioning with attention



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

Image Captioning with Attention



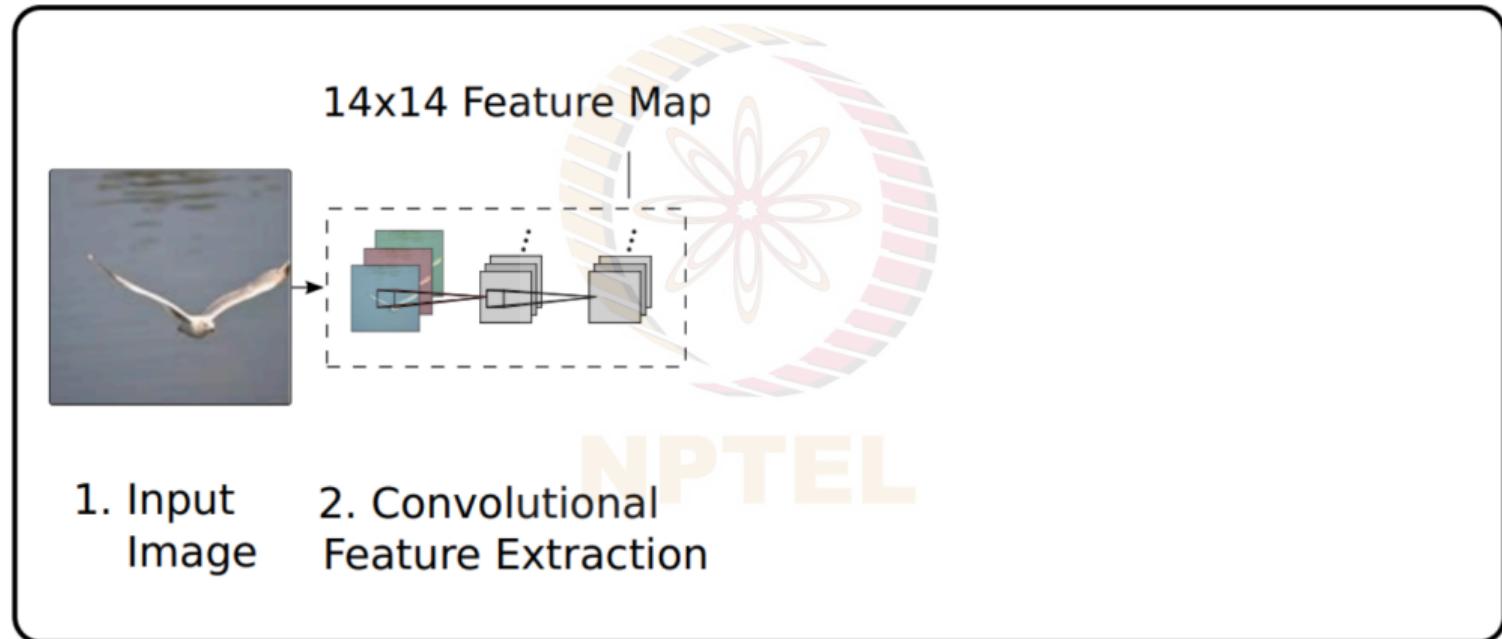
1. Input
Image



NPTEL

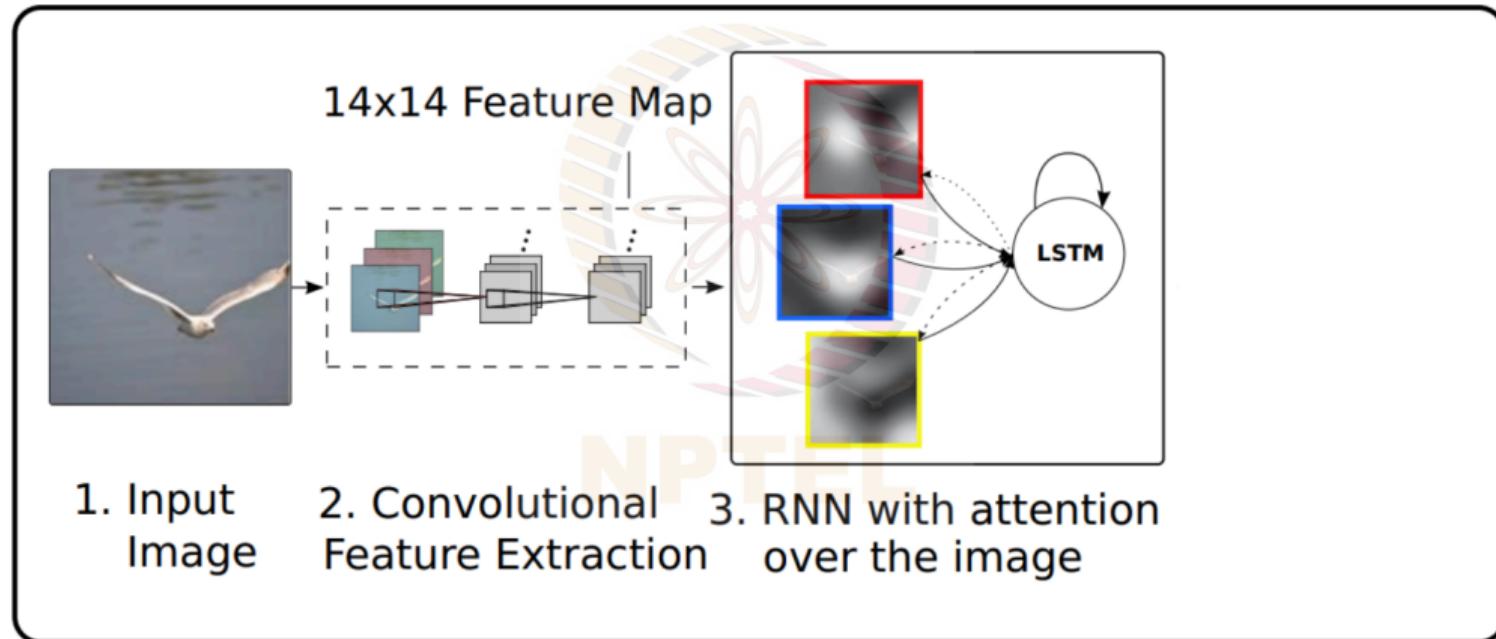
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



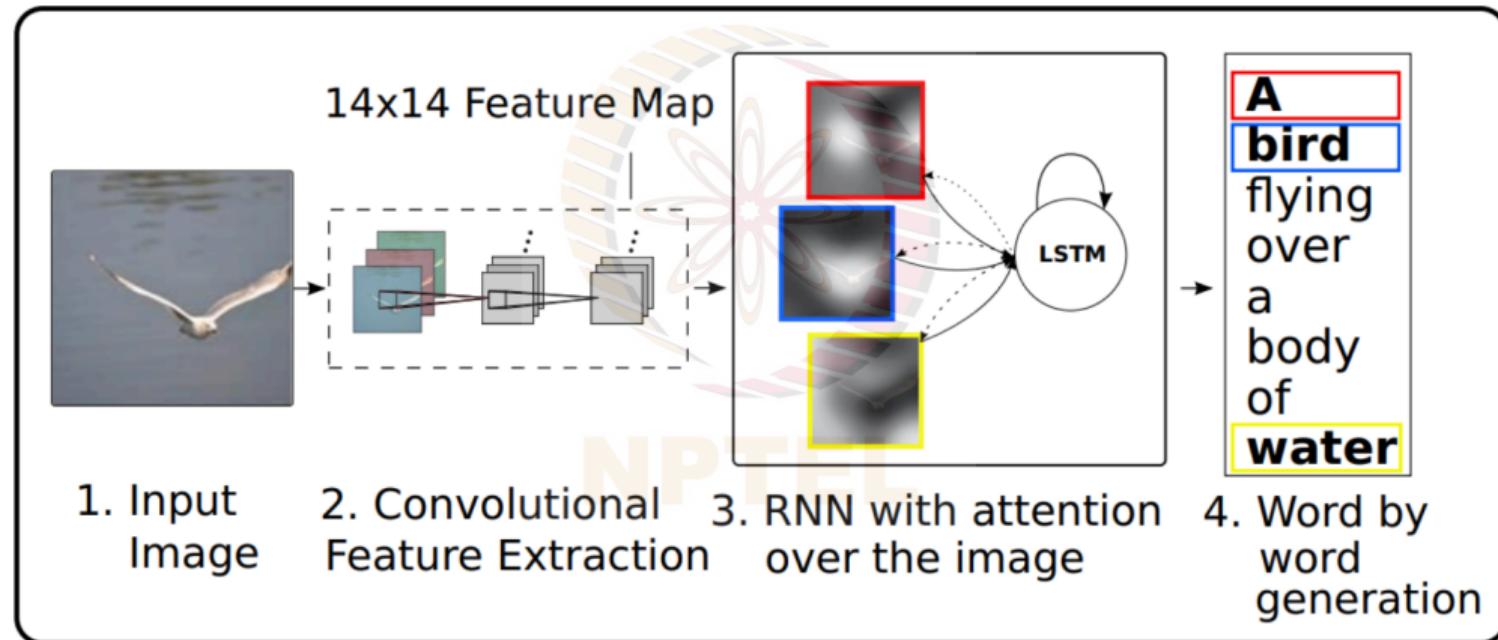
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



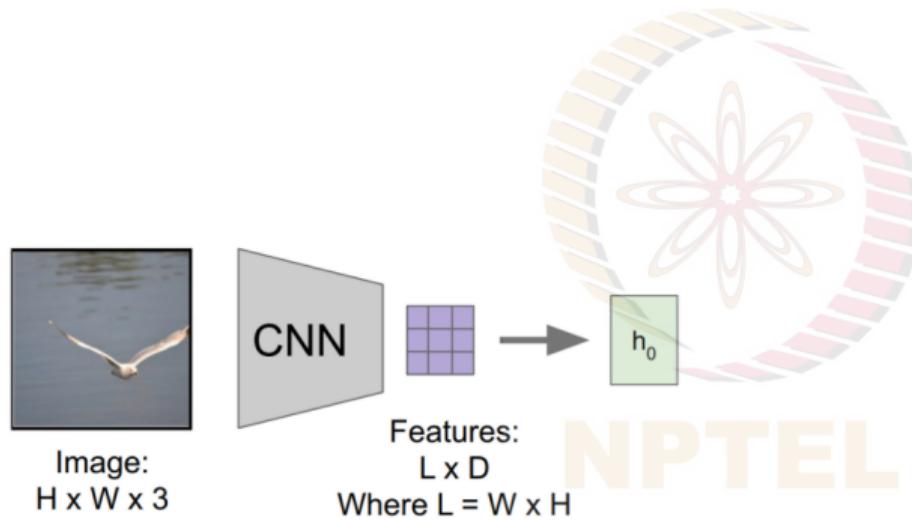
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



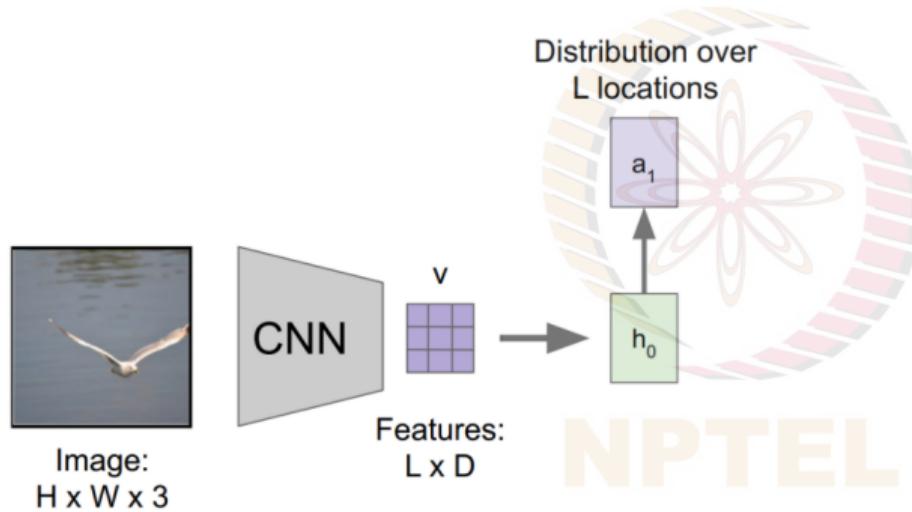
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



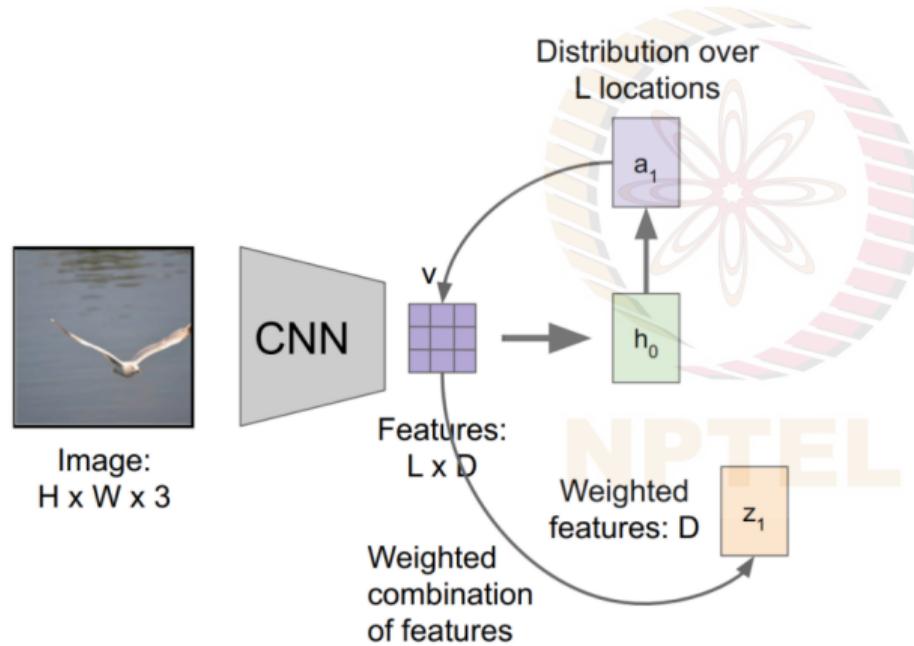
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



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

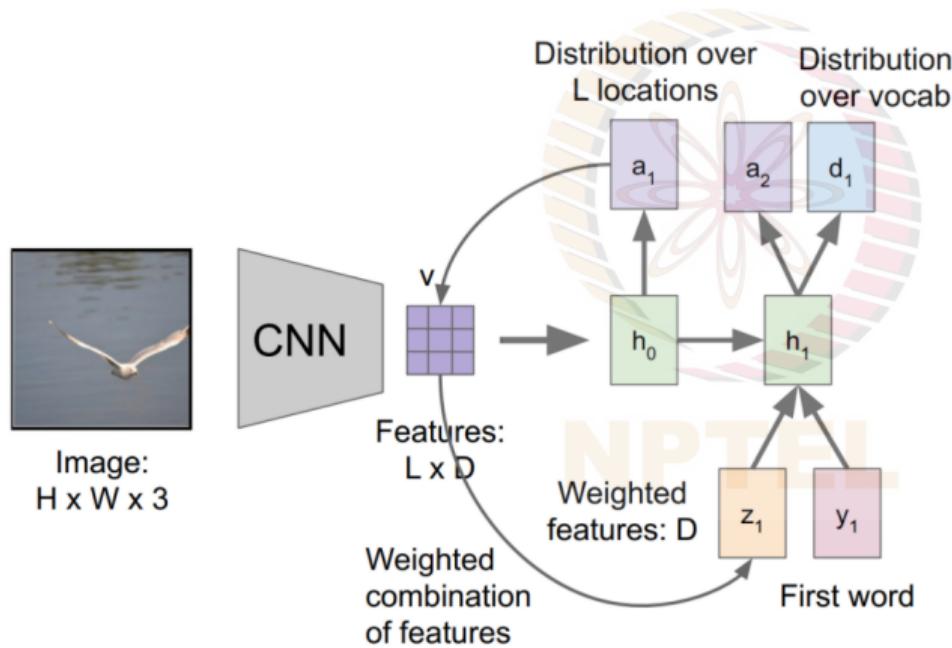
Image Captioning with Attention



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

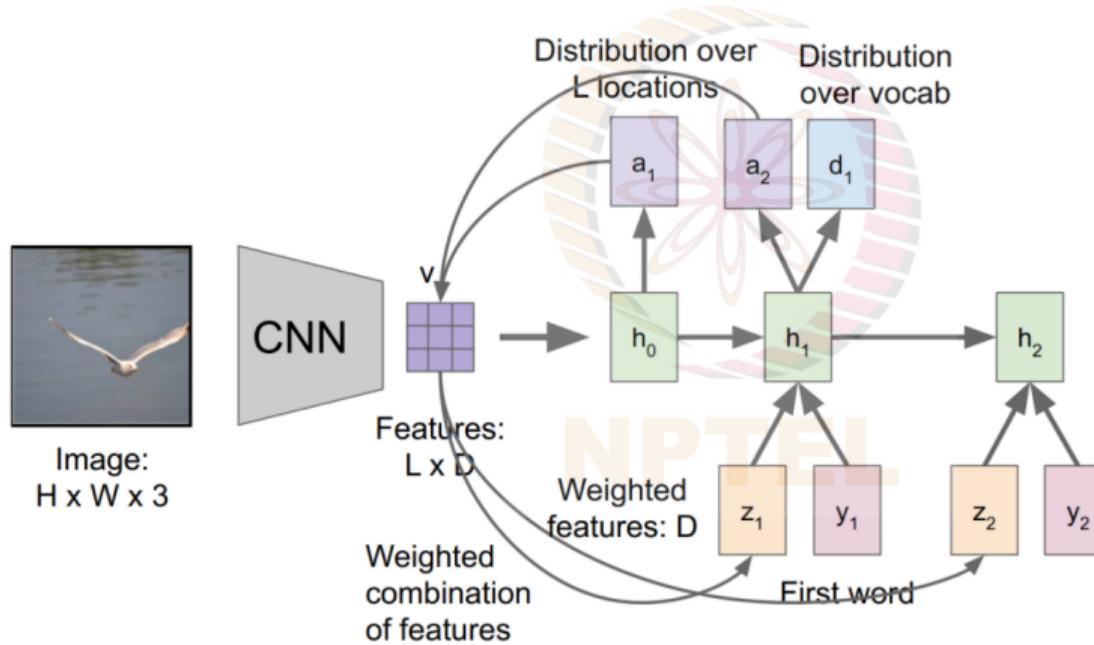
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



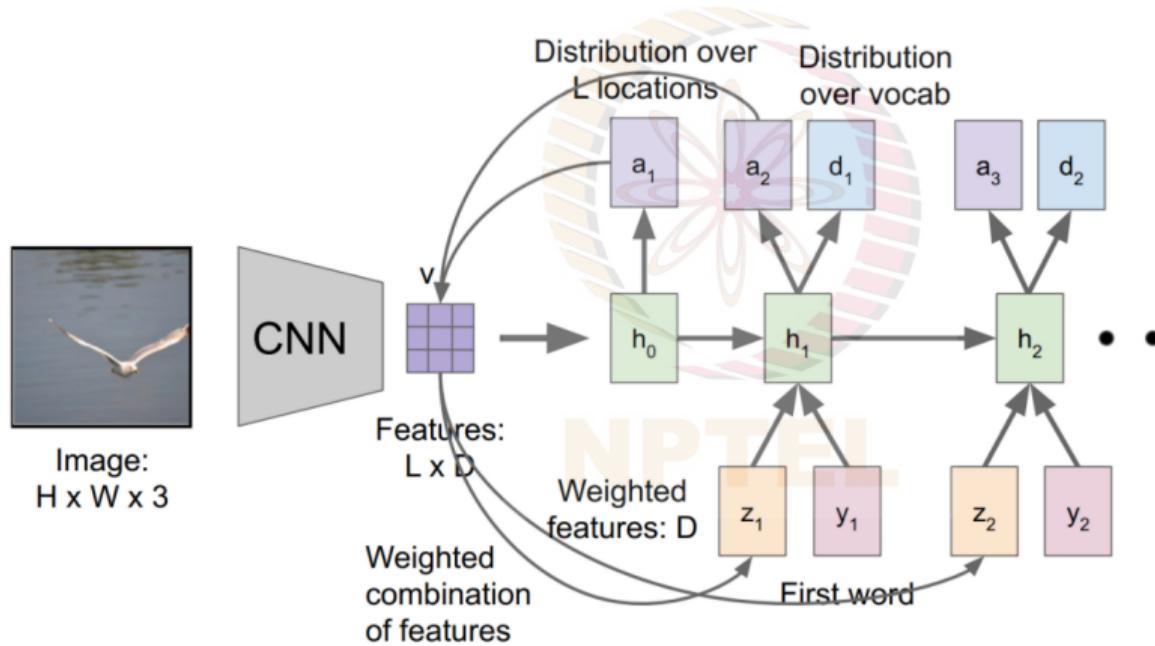
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



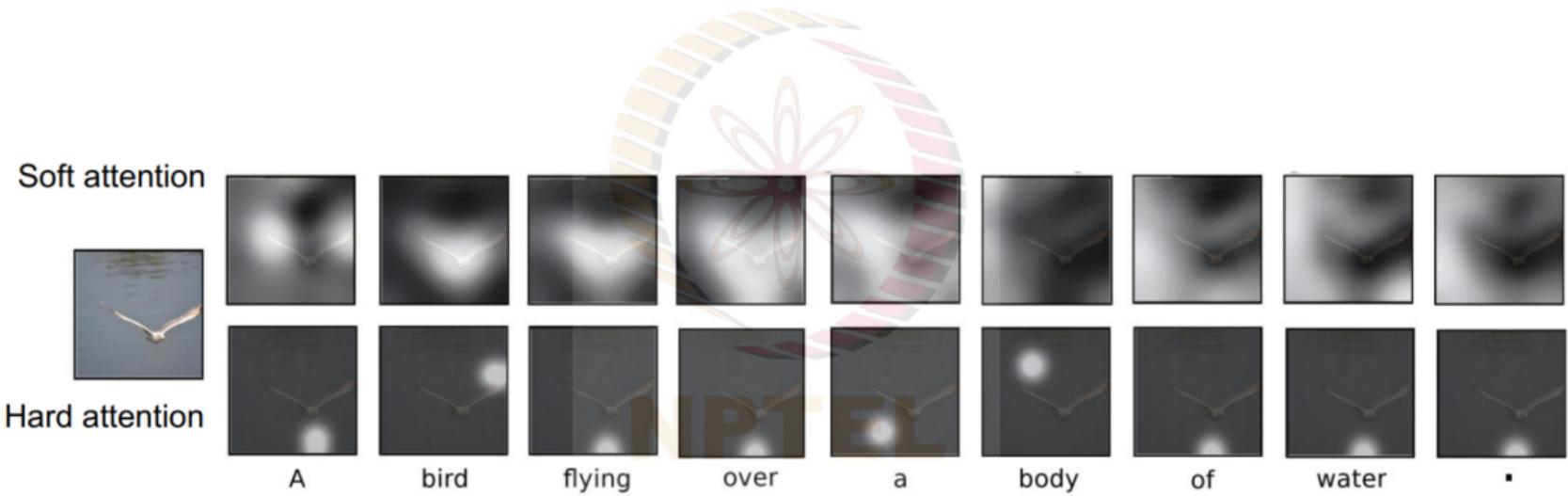
Credit: Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Image Captioning with Attention



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.



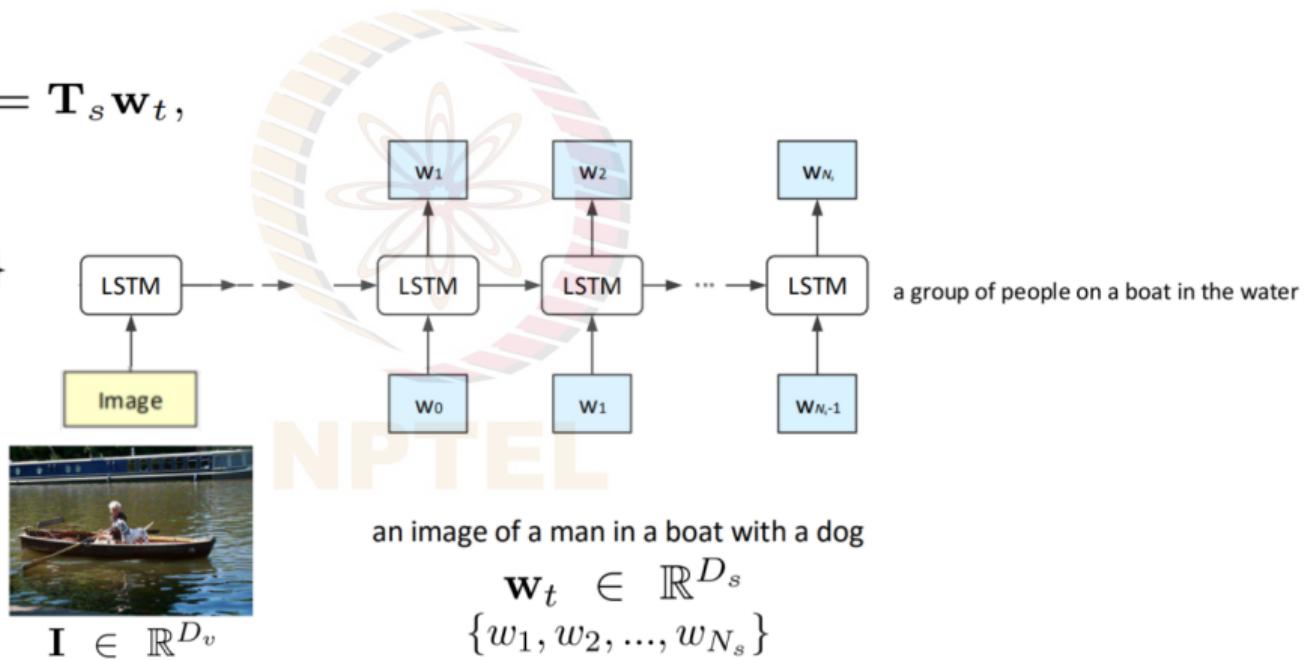
A giraffe standing in a forest with trees in the background.

Recent Efforts: Boosting Image Captioning with Attributes in LSTMs

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

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

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

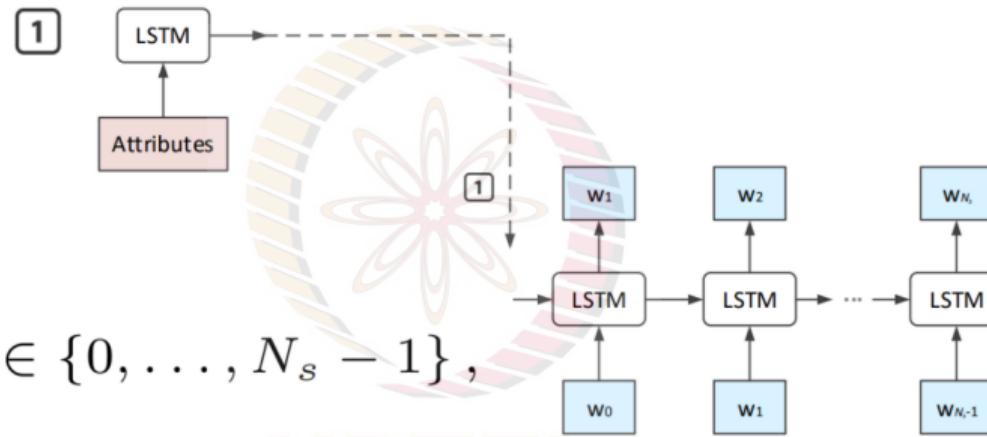


Boosting Image Captioning with Attributes in LSTMs: A1

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

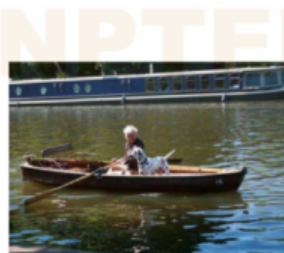
$$\mathbf{x}^t = \mathbf{T}_s \mathbf{w}_t,$$

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



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

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



Attributes:

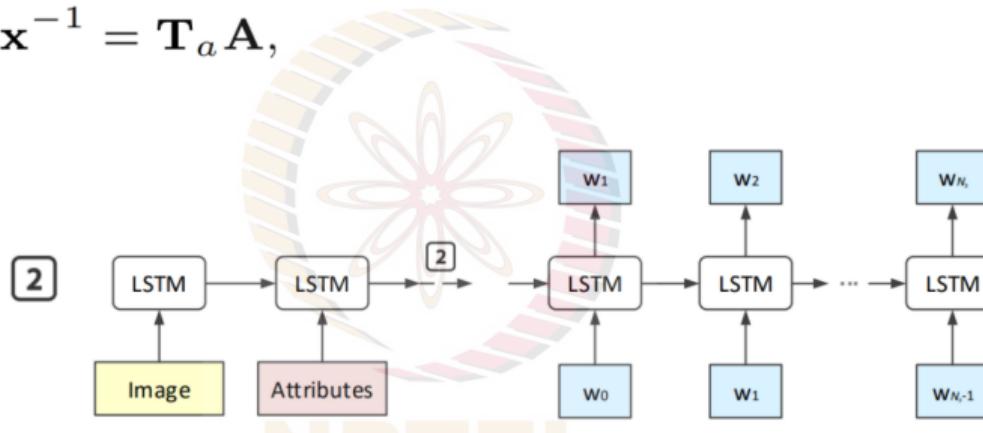
boat: 1 water: 0.838 man: 0.762
riding: 0.728 dog: 0.547 small: 0.485
person: 0.471 river: 0.461

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

Boosting Image Captioning with Attributes in LSTMs: A2

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

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



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

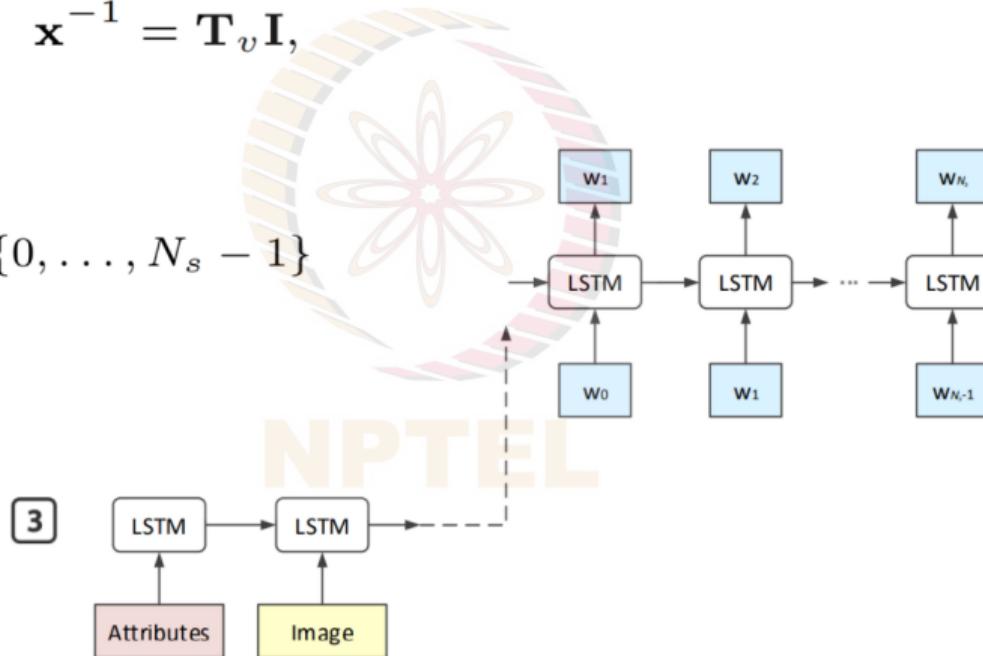
Credit: Yao et al, Boosting Image Captioning with Attributes, ICCV 2017

Boosting Image Captioning with Attributes in LSTMs: A3

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

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

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



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

Boosting Image Captioning with Attributes in LSTMs: A4

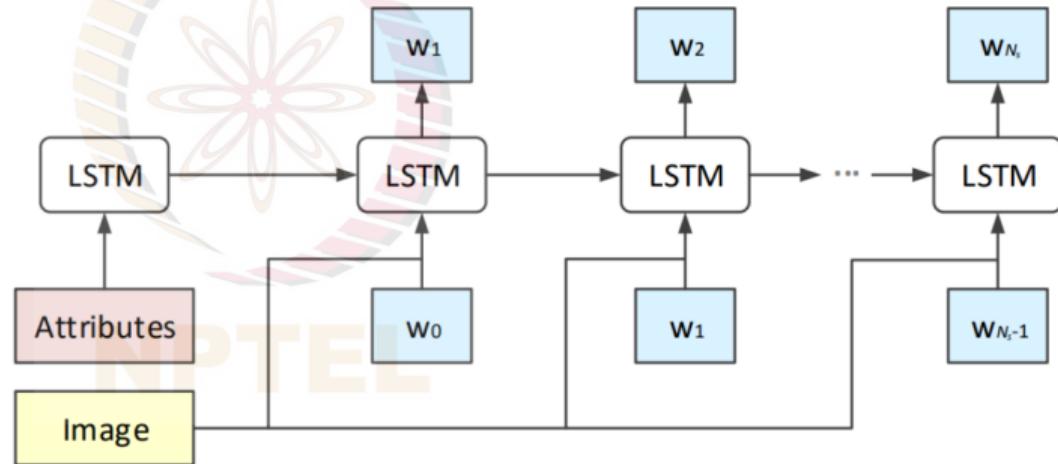
$$\mathbf{x}^{-1} = \mathbf{T}_a \mathbf{A},$$

4

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

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

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



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

Boosting Image Captioning with Attributes in LSTMs: A5

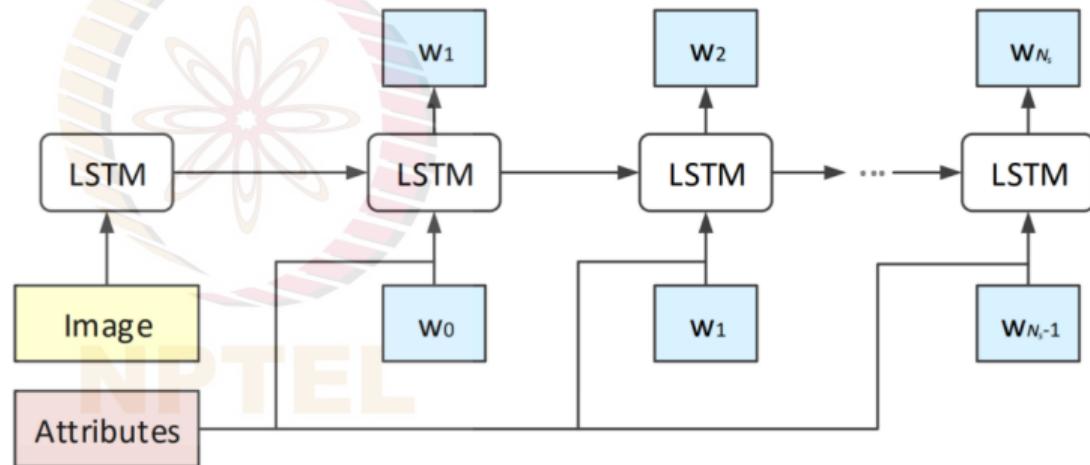
$$\mathbf{x}^{-1} = \mathbf{T}_v \mathbf{I},$$

5

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

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

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



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

Boosting Image Captioning with Attributes in LSTMs: Observations

- $\text{LSTM-}A_1 > \text{LSTM}$
 - Indicates advantage of exploiting high-level attributes than image representations



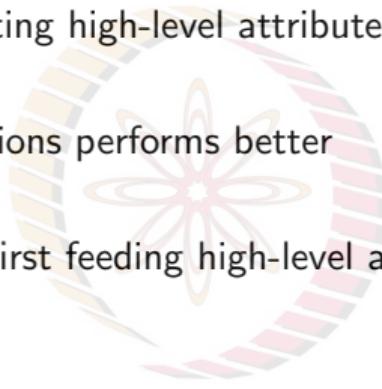
Boosting Image Captioning with Attributes in LSTMs: Observations

- $\text{LSTM-}A_1 > \text{LSTM}$
 - Indicates advantage of exploiting high-level attributes than image representations
- $\text{LSTM-}A_2 > \text{LSTM-}A_1$
 - Integrating image representations performs better



Boosting Image Captioning with Attributes in LSTMs: Observations

- $\text{LSTM-}A_1 > \text{LSTM}$
 - Indicates advantage of exploiting high-level attributes than image representations
- $\text{LSTM-}A_2 > \text{LSTM-}A_1$
 - Integrating image representations performs better
- $\text{LSTM-}A_3 > \text{LSTM-}A_2$
 - Benefits from mechanism of first feeding high-level attributes into LSTM instead of starting from image representations



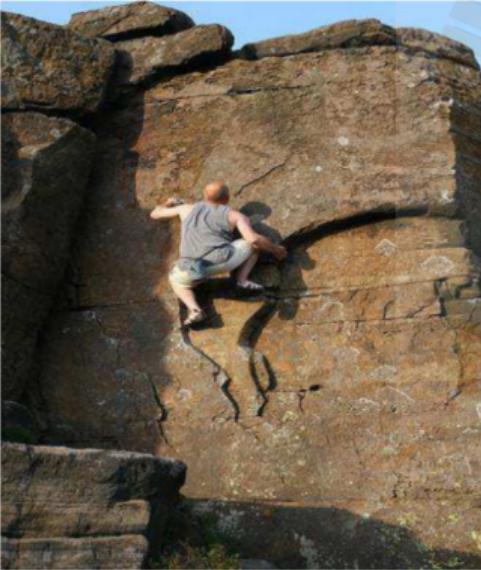
Boosting Image Captioning with Attributes in LSTMs: Observations

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

Boosting Image Captioning with Attributes in LSTMs: Observations

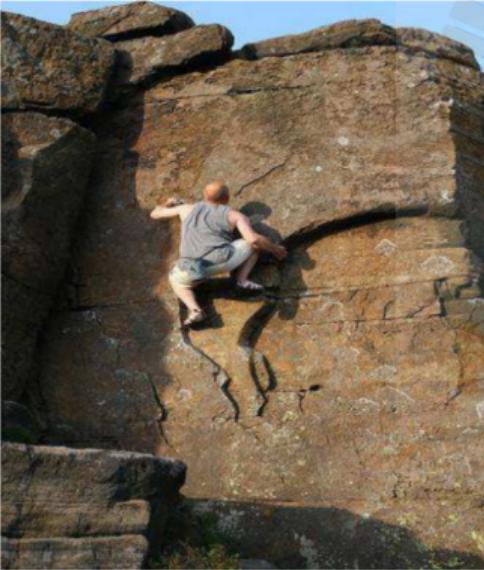
- $\text{LSTM-}A_1 > \text{LSTM}$
 - Indicates advantage of exploiting high-level attributes than image representations
- $\text{LSTM-}A_2 > \text{LSTM-}A_1$
 - Integrating image representations performs better
- $\text{LSTM-}A_3 > \text{LSTM-}A_2$
 - Benefits from mechanism of first feeding high-level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

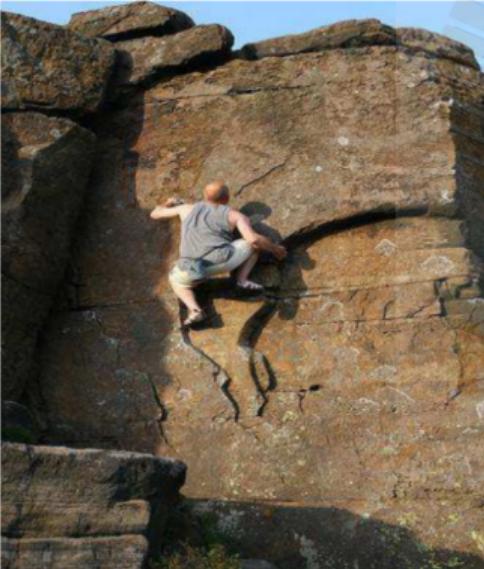
StyleNet: Generating Attractive Visual Captions with Styles



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

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

StyleNet: Generating Attractive Visual Captions with Styles



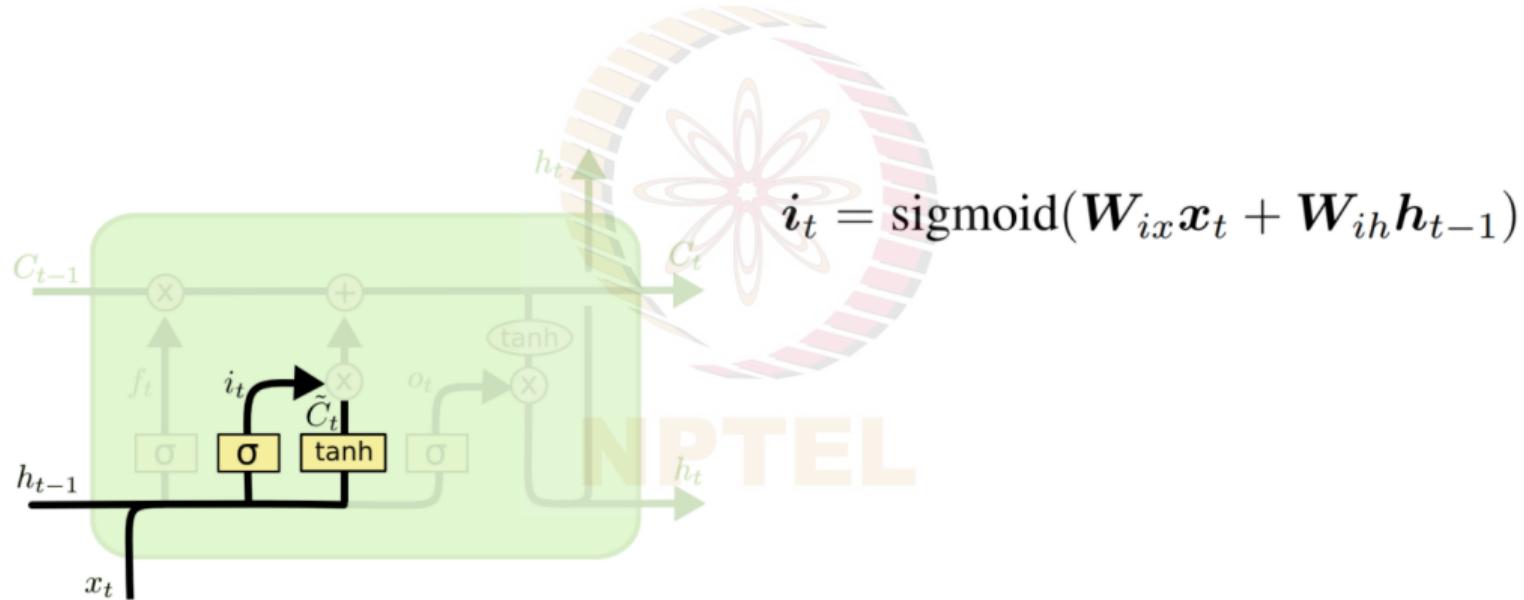
CaptionBot: A man on a rocky hillside
next to a stone wall.

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

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

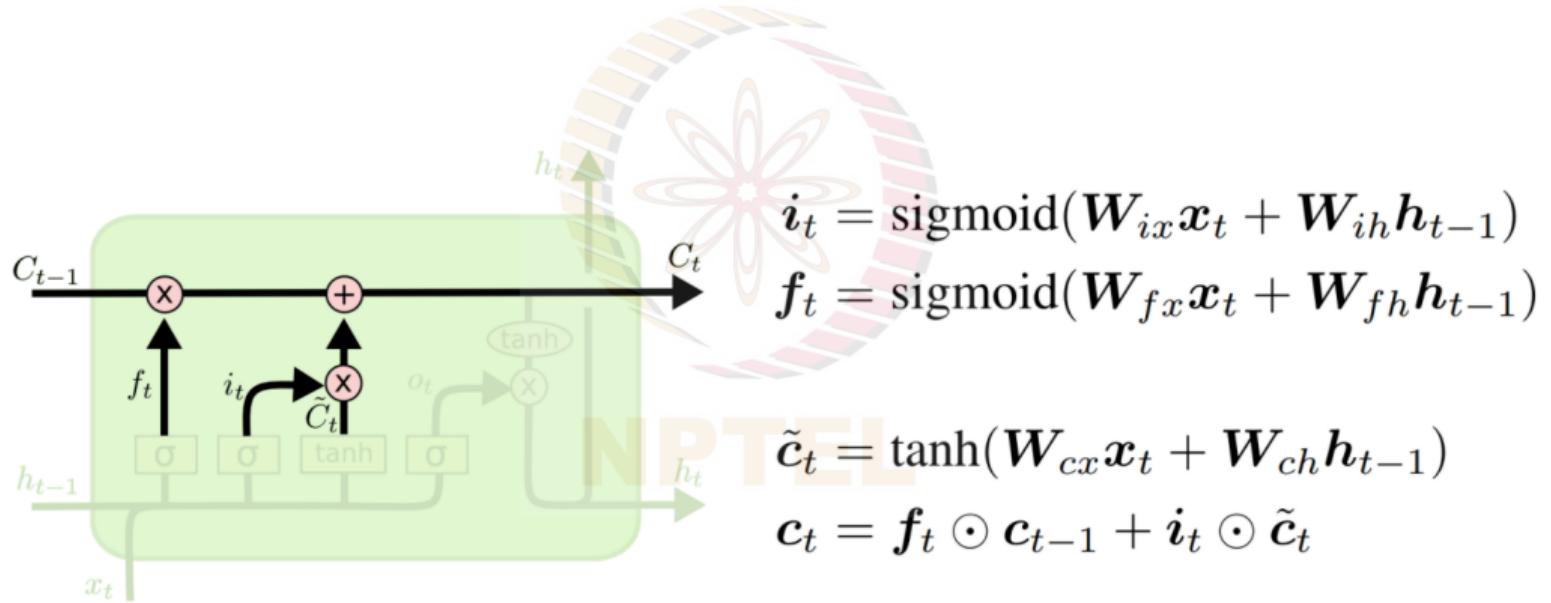
StyleNet: Generating Attractive Visual Captions with Styles

Recall LSTM



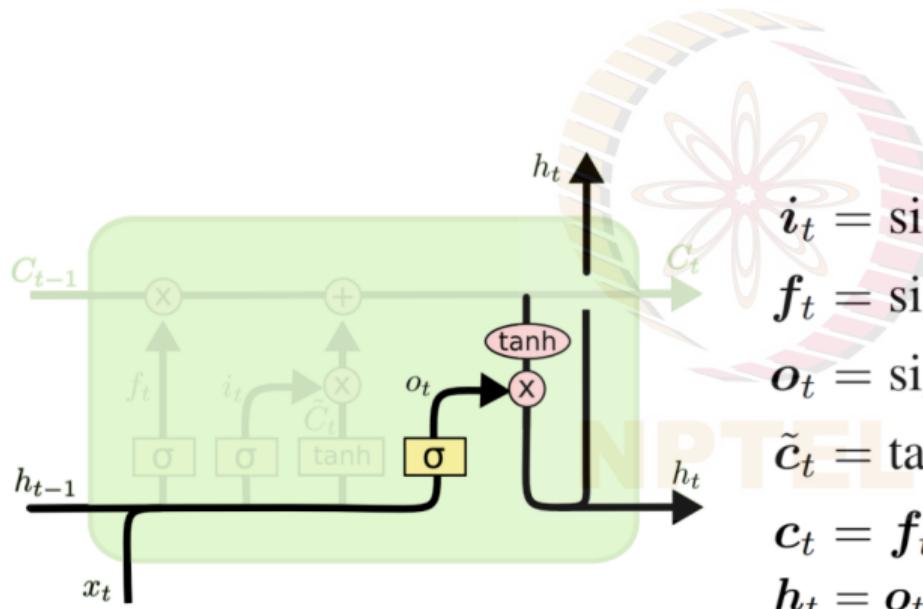
StyleNet: Generating Attractive Visual Captions with Styles

Recall LSTM



StyleNet: Generating Attractive Visual Captions with Styles

Recall LSTM



$$i_t = \text{sigmoid}(\mathbf{W}_{ix}x_t + \mathbf{W}_{ih}h_{t-1})$$

$$f_t = \text{sigmoid}(\mathbf{W}_{fx}x_t + \mathbf{W}_{fh}h_{t-1})$$

$$o_t = \text{sigmoid}(\mathbf{W}_{ox}x_t + \mathbf{W}_{oh}h_{t-1})$$

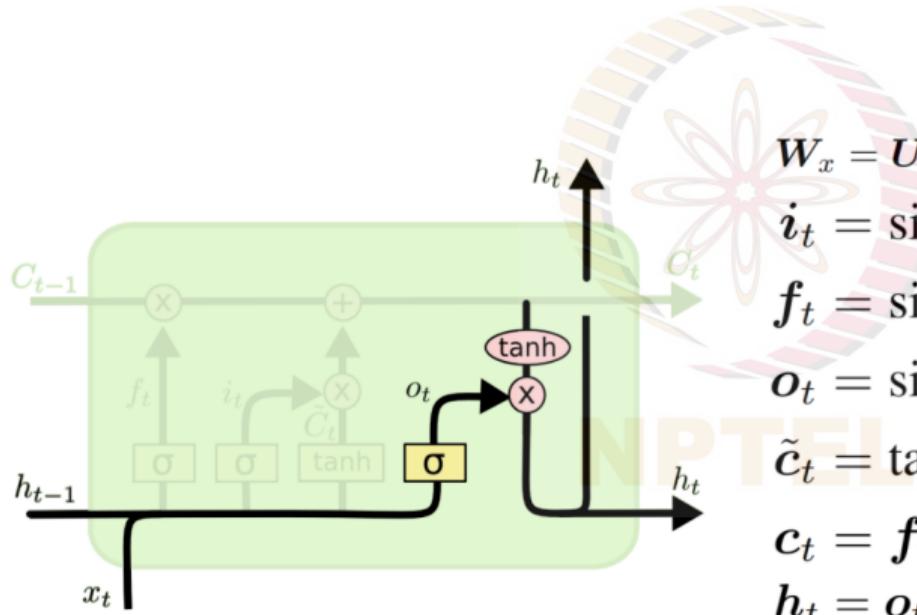
$$\tilde{c}_t = \tanh(\mathbf{W}_{cx}x_t + \mathbf{W}_{ch}h_{t-1})$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$h_t = o_t \odot c_t$$

StyleNet: Generating Attractive Visual Captions with Styles

StyleNet proposes a factored LSTM



$$\mathbf{W}_x = \mathbf{U}_x \mathbf{S}_x \mathbf{V}_x$$

$$i_t = \text{sigmoid}(\mathbf{W}_{ix} \mathbf{x}_t + \mathbf{W}_{ih} \mathbf{h}_{t-1})$$

$$f_t = \text{sigmoid}(\mathbf{W}_{fx} \mathbf{x}_t + \mathbf{W}_{fh} \mathbf{h}_{t-1})$$

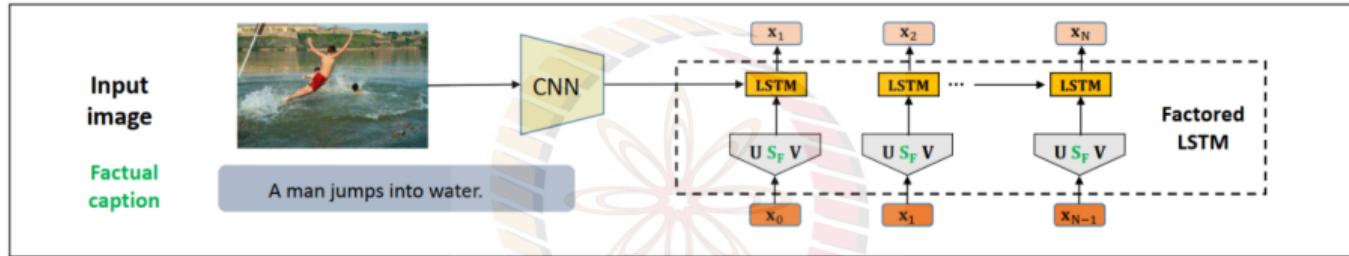
$$o_t = \text{sigmoid}(\mathbf{W}_{ox} \mathbf{x}_t + \mathbf{W}_{oh} \mathbf{h}_{t-1})$$

$$\tilde{c}_t = \tanh(\mathbf{W}_{cx} \mathbf{x}_t + \mathbf{W}_{ch} \mathbf{h}_{t-1})$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$h_t = o_t \odot c_t$$

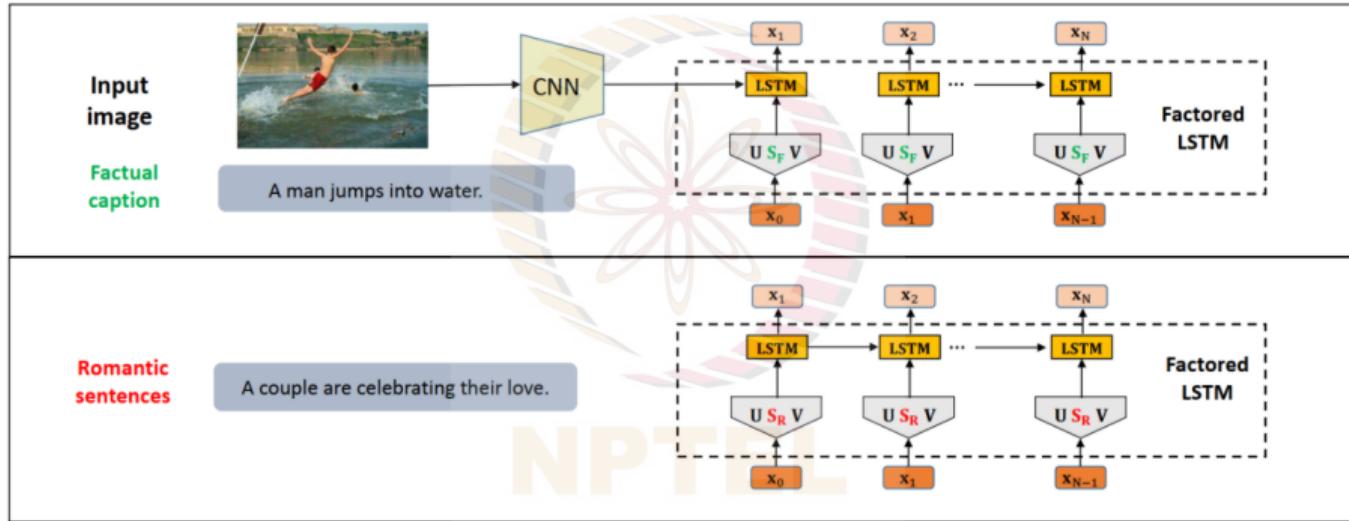
StyleNet: Generating Attractive Visual Captions with Styles



NPTEL

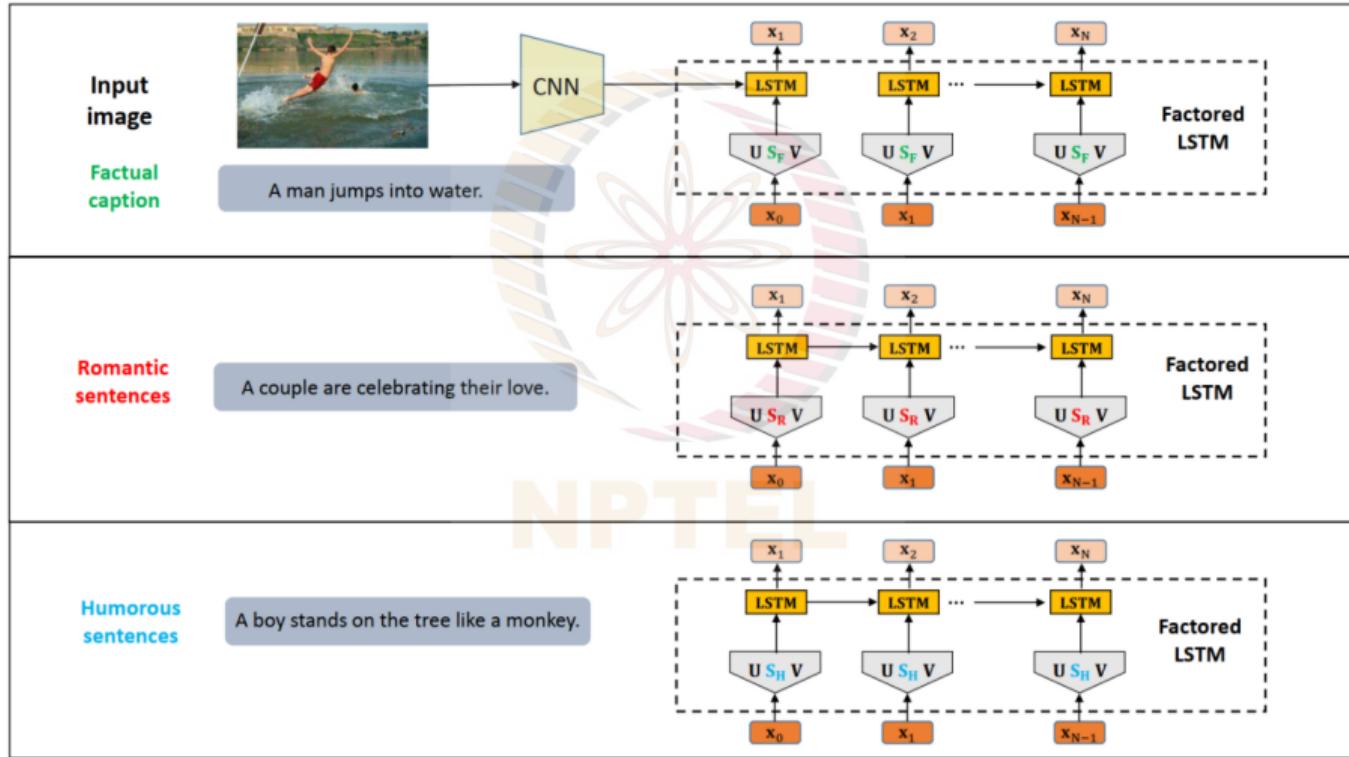
U, S_F, V are trainable.

StyleNet: Generating Attractive Visual Captions with Styles



Keep U, V same as earlier.
 S_R is trainable.

StyleNet: Generating Attractive Visual Captions with Styles



StyleNet: Generating Attractive Visual Captions with Styles



F: A snowboarder in the air .

R: A man is doing a trick on a skateboard to show his courage .

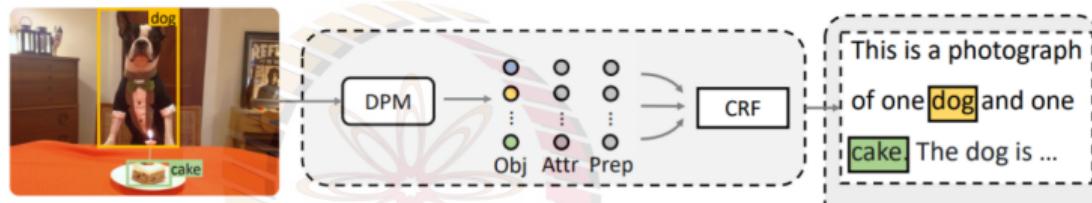
H: A man is jumping on a snowboard to reach outer space .

NPTEL

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

Neural Baby Talk

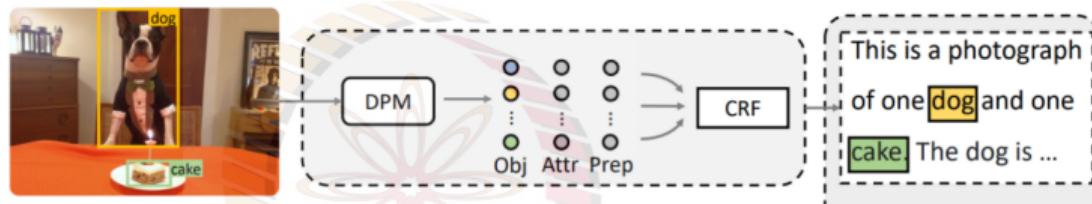
More Grounded



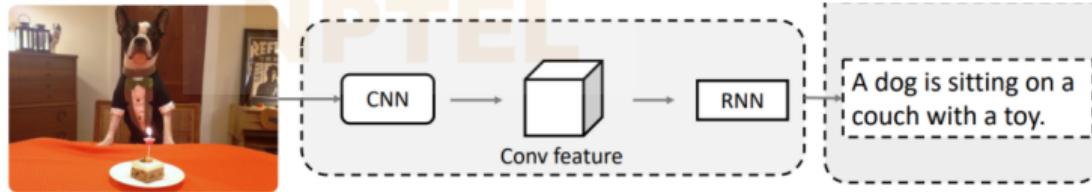
Credit: Lu et al, Neural Baby Talk, CVPR 2018

Neural Baby Talk

More Grounded



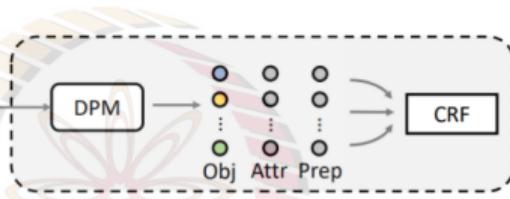
More Natural



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

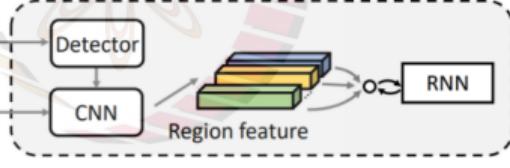
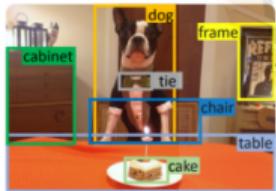
Neural Baby Talk

More Grounded



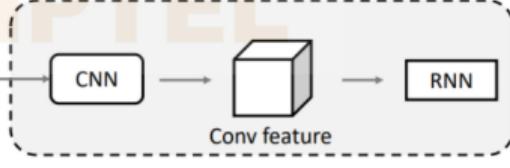
This is a photograph
of one **dog** and one
cake. The dog is ...

Neural Baby Talk



A (**yellow**) with a (**grey**) is
sitting at (**blue**) with
a (**green**).
[Legend:
Yellow box → puppy
Grey box → tie
Green box → cake
Blue box → table]

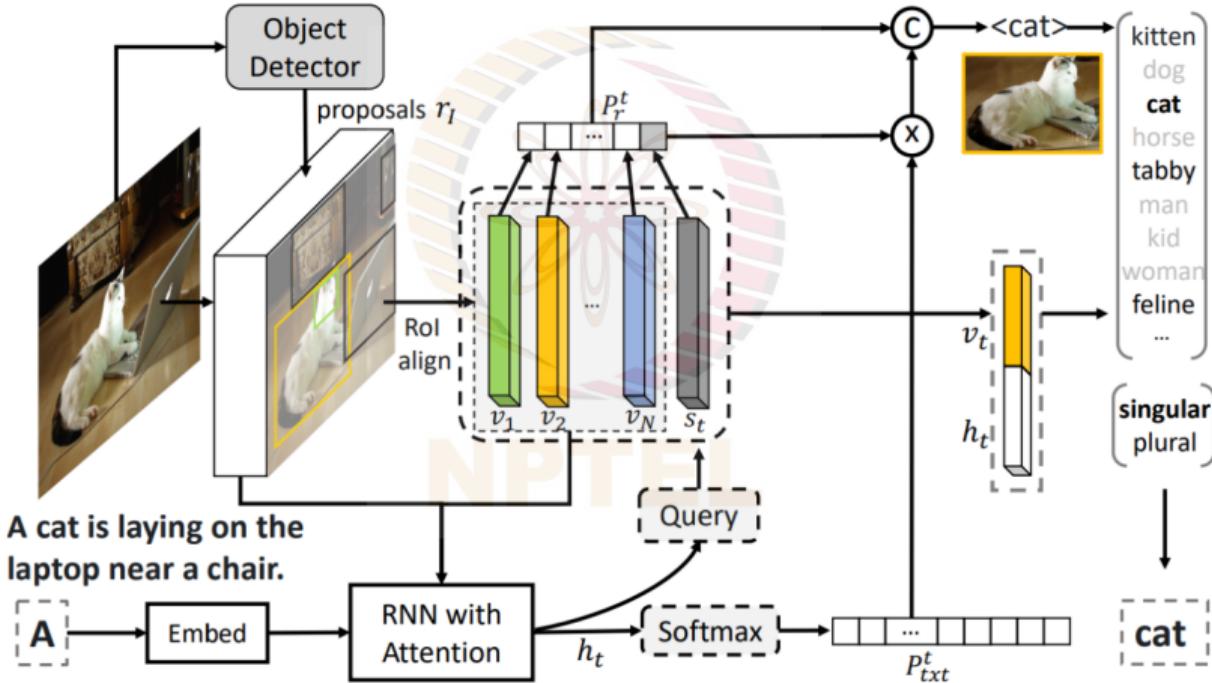
More Natural



A dog is sitting on a
couch with a toy.

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

Neural Baby Talk



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

Homework

Readings

- Papers on respective slides
- (Optional) [GeorgiaTech Notebook on Image Captioning \(Part: 1-3\)](#)



Question

- Can we do the opposite (caption-to-image) of what we learnt in this section? How?