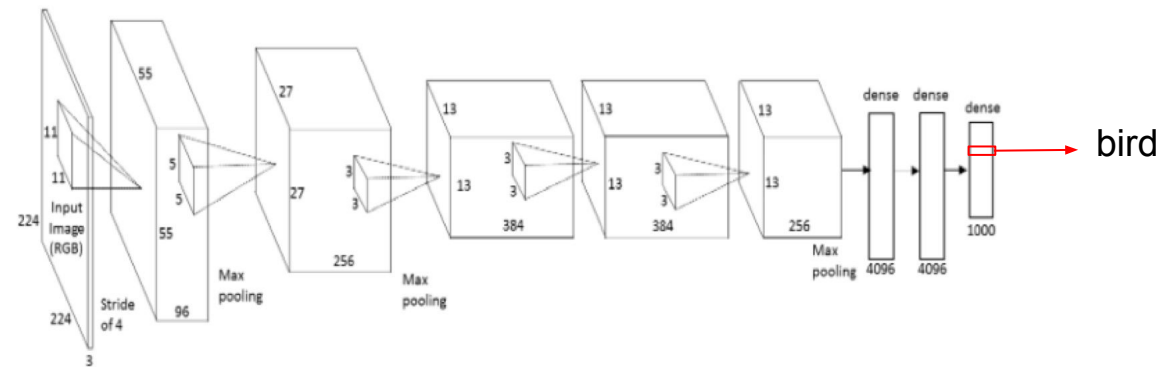


Attention Models: Motivation



Image:
 $H \times W \times 3$



The whole input volume is used to predict the output...

Attention Models: Motivation

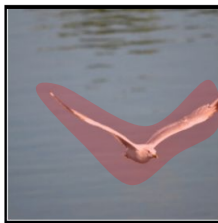
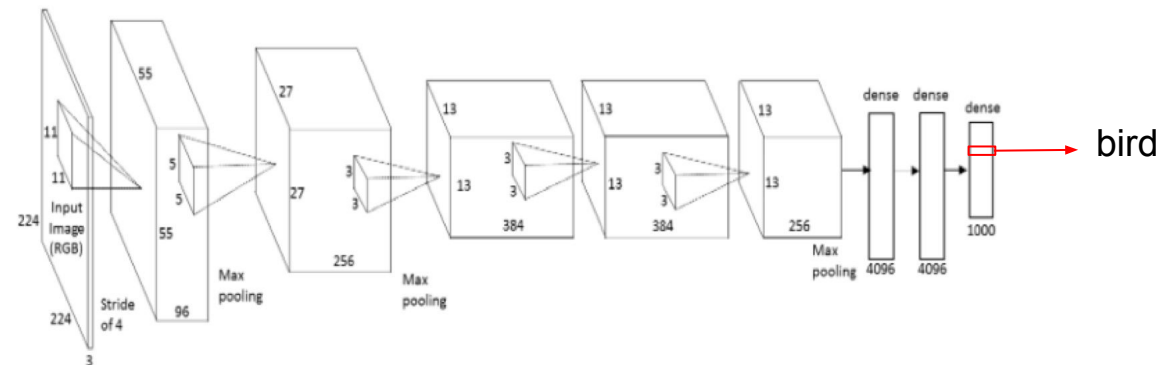


Image:
 $H \times W \times 3$



The whole input volume is used to predict the output...

...despite the fact that not all pixels are equally important

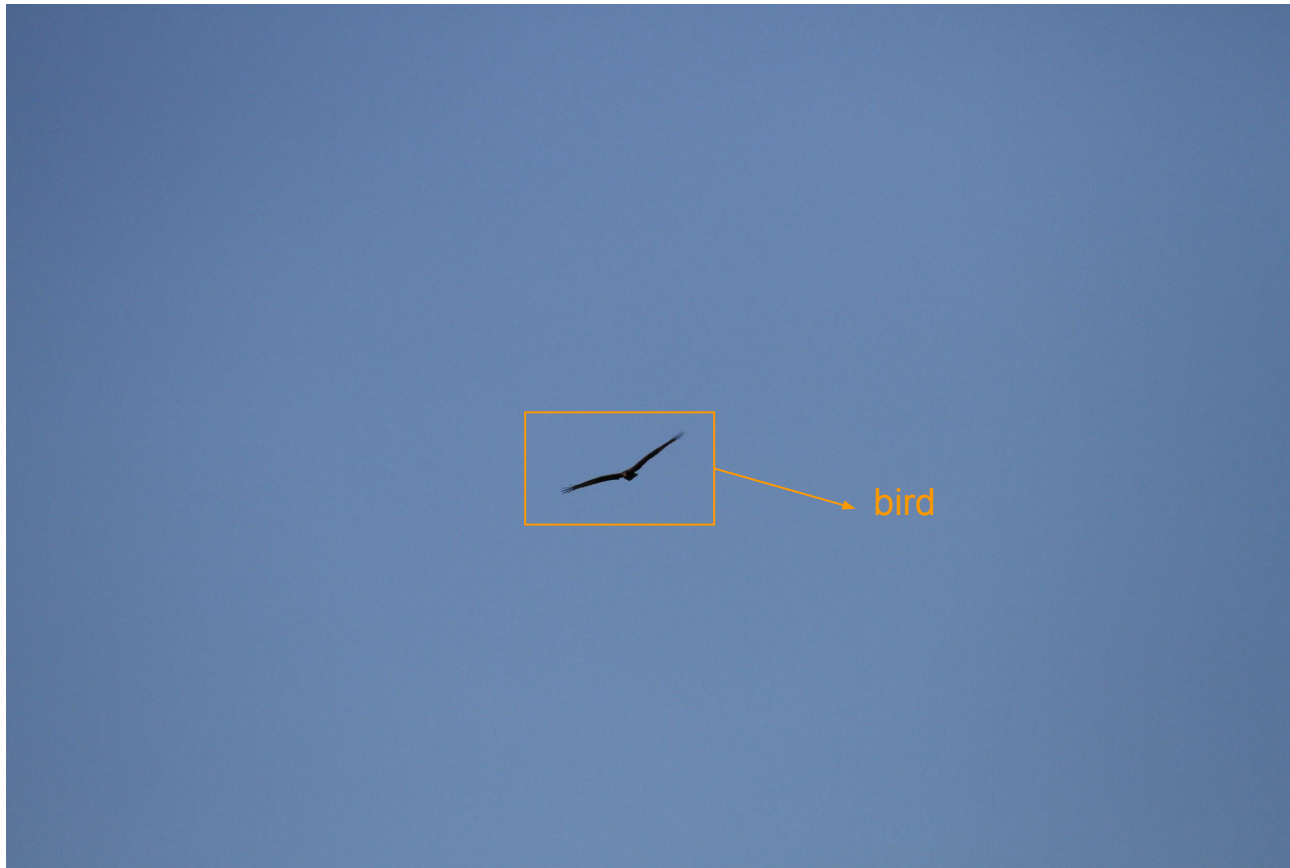
Attention Models: Motivation



Attention models can
relieve computational burden

Helpful when processing big
images !

Attention Models: Motivation



Attention models can
relieve computational burden

Helpful when processing big
images !

Attention Models

Attend to different parts of the input to optimize a certain output

Attention Models

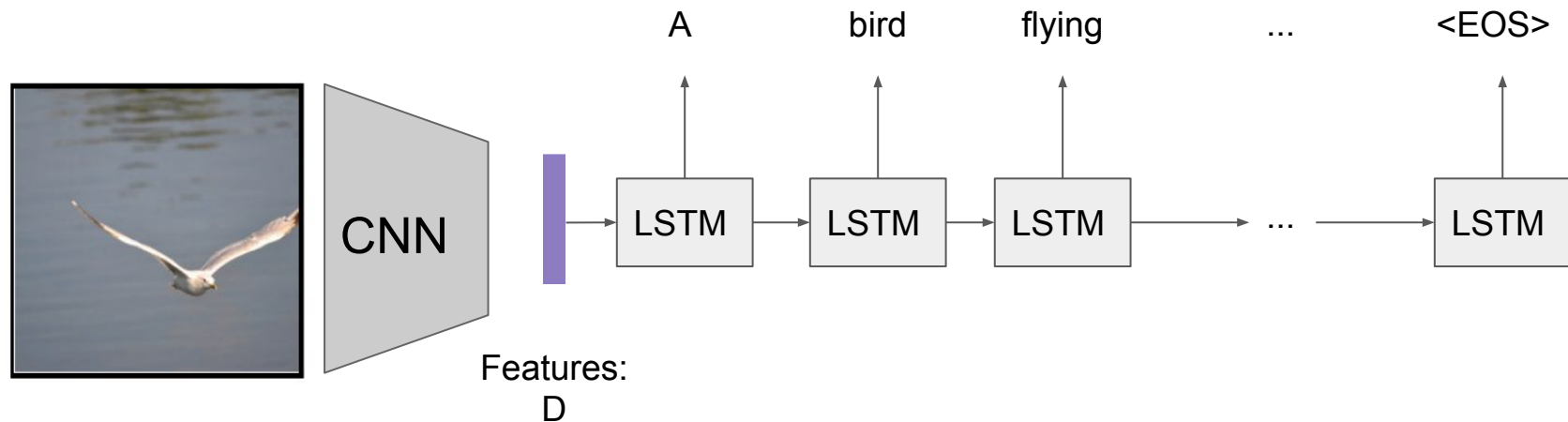
Attend to different parts of the input to optimize a certain output

Input: Image; Output: Text



A bird flying over a body of water

LSTM Decoder for Image Captioning



The LSTM decoder “sees” the input only at the beginning !

Vinyals et al. [Show and tell: A neural image caption generator.](#) CVPR 2015

Attention for Image Captioning

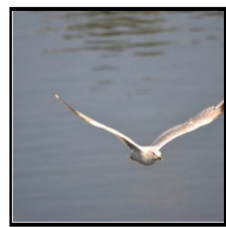
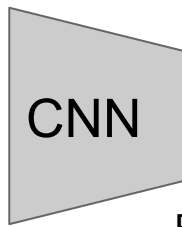
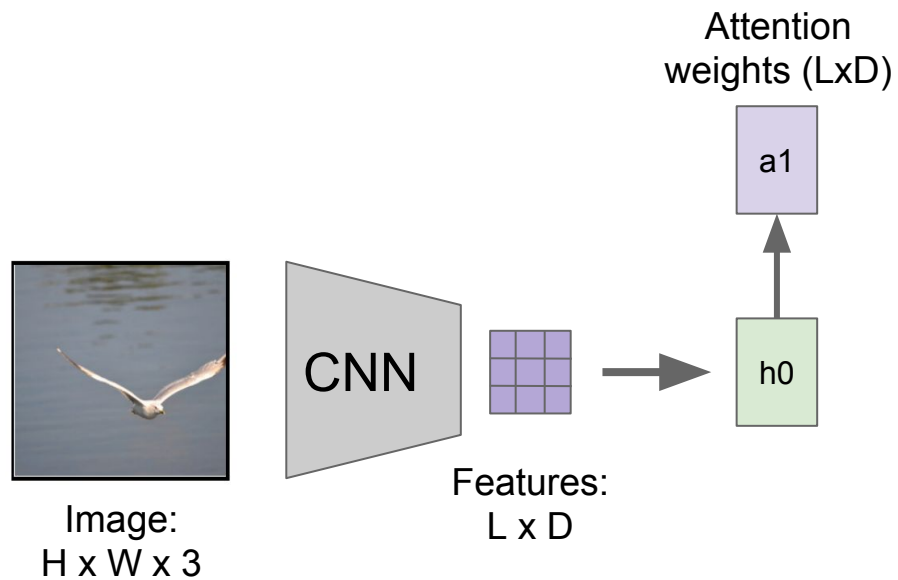


Image:
 $H \times W \times 3$

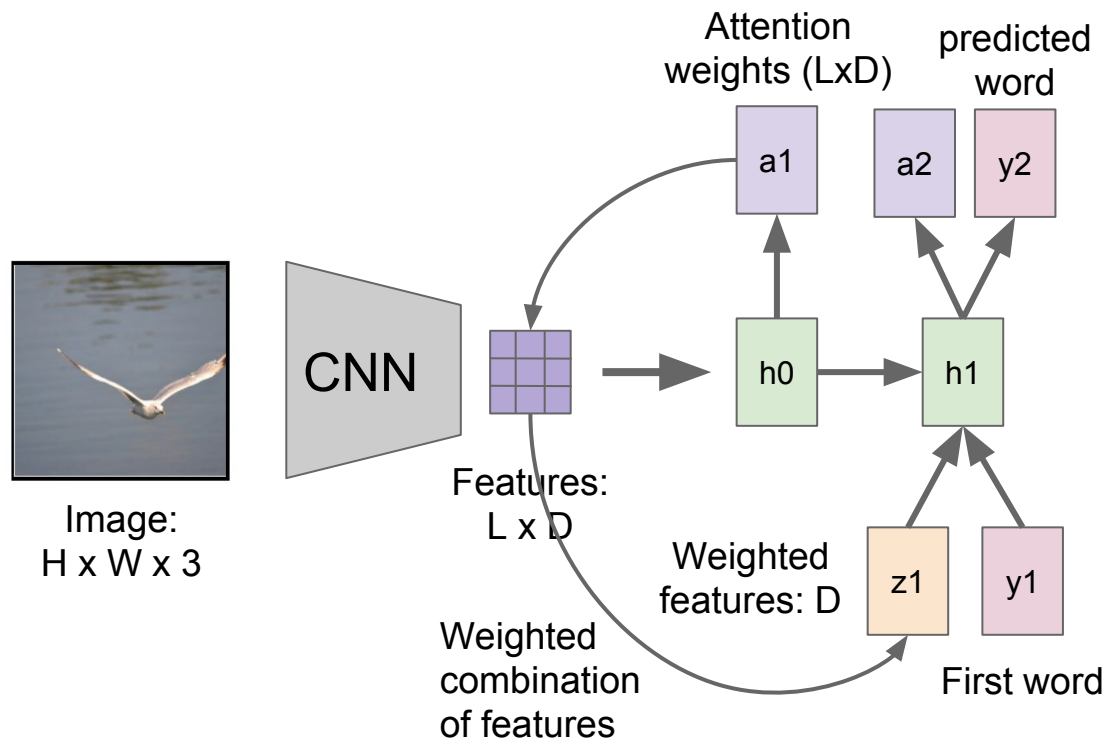


Features:
 $L \times D$

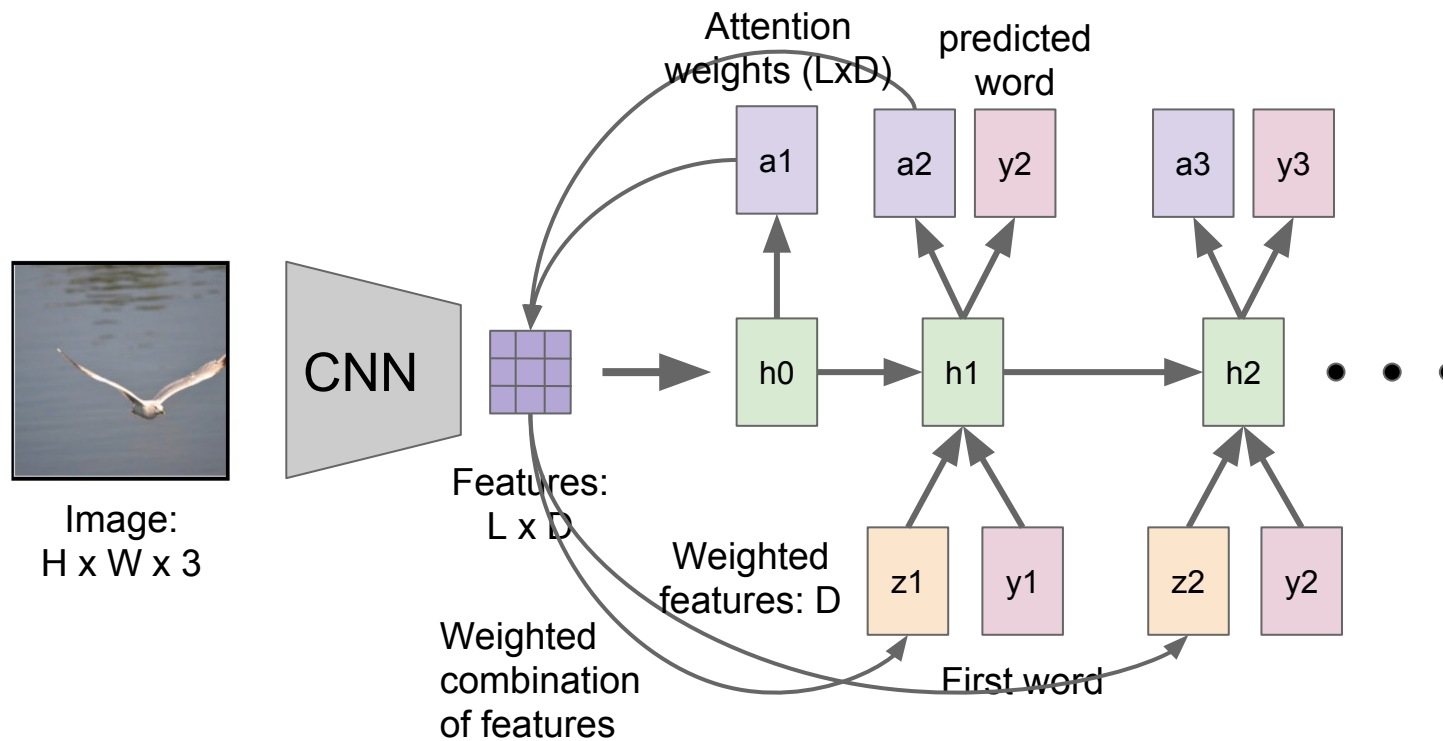
Attention for Image Captioning



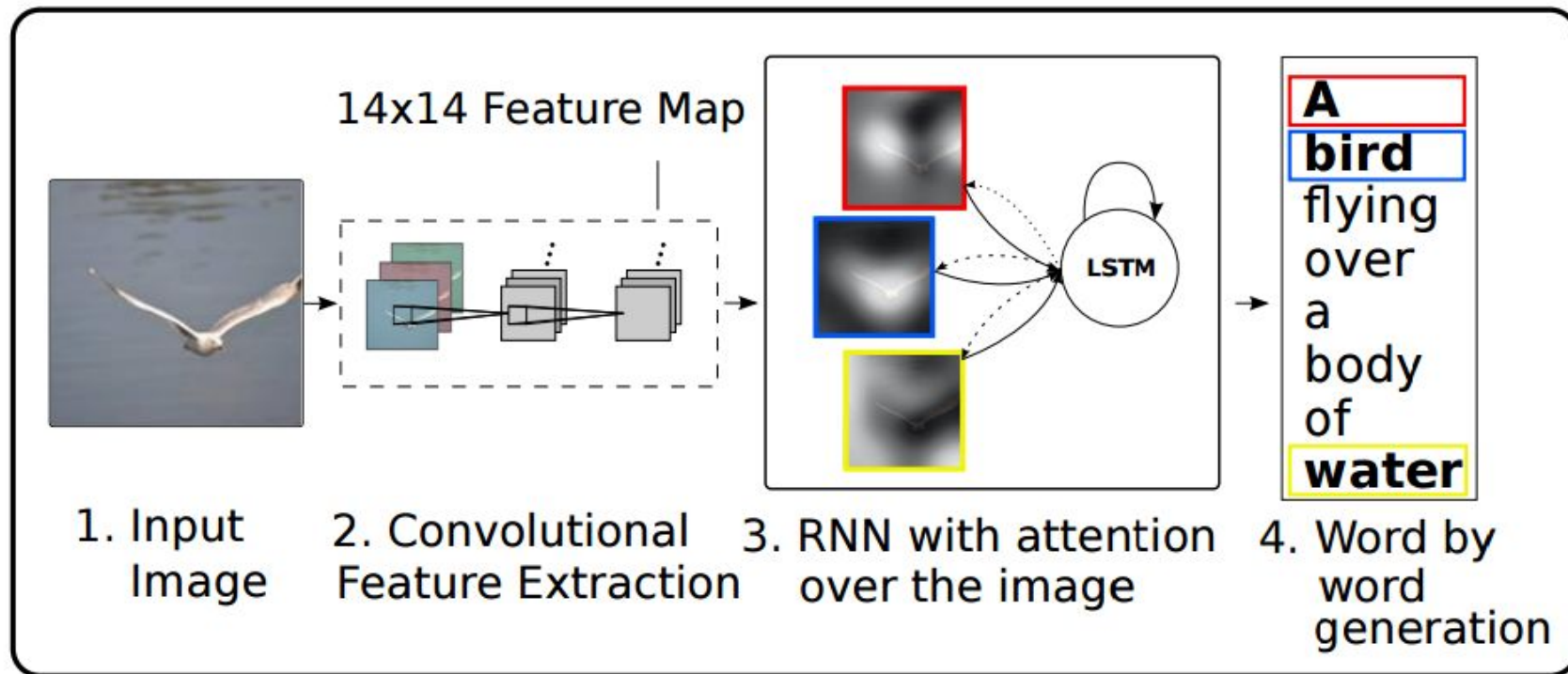
Attention for Image Captioning



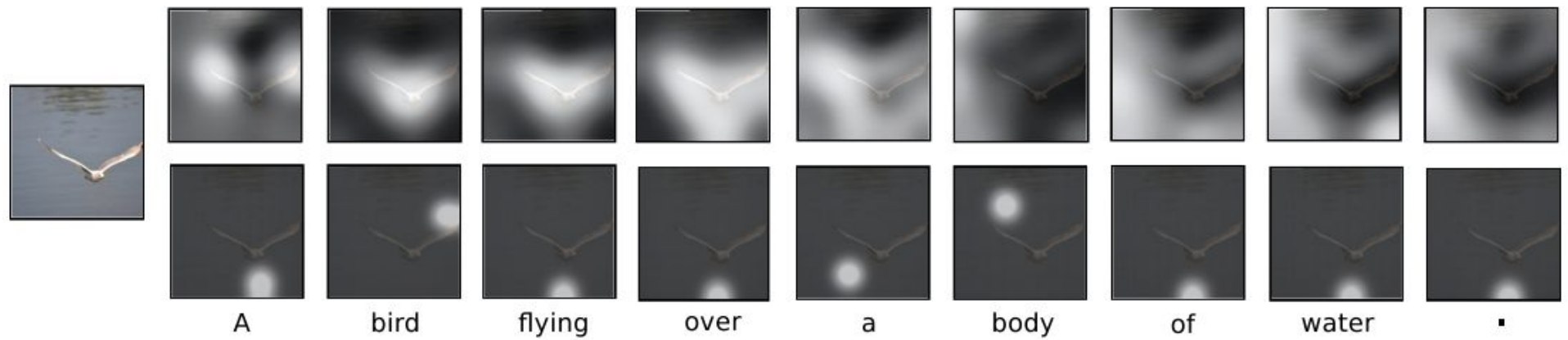
Attention for Image Captioning



Attention for Image Captioning



Attention for Image Captioning



Xu et al. [Show, Attend and Tell: Neural Image Caption Generation with Visual Attention](#). ICML 2015

Attention for Image Captioning



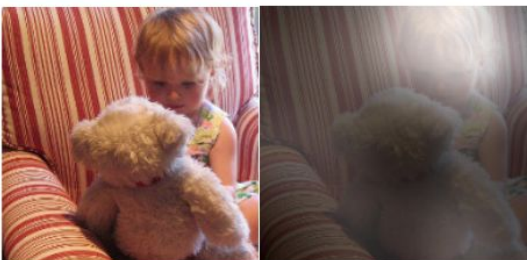
A woman is throwing a frisbee in a park.



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.

Xu et al. [Show, Attend and Tell: Neural Image Caption Generation with Visual Attention](#). ICML 2015