Attention Models

Jaegul Choo (주재걸)
Korea University
https://sites.google.com/site/jaegulchoo/

Contents

1. Recurrent Neural Network

1-1. RNN의 여러가지 형태

1-2. Character-level LM

1-3. Vanishing Gradient Problem

2. LSTM과 GRU

2-1. LSTM이란 무엇인가

2-2. GRU란 무엇인가

Image Captioning

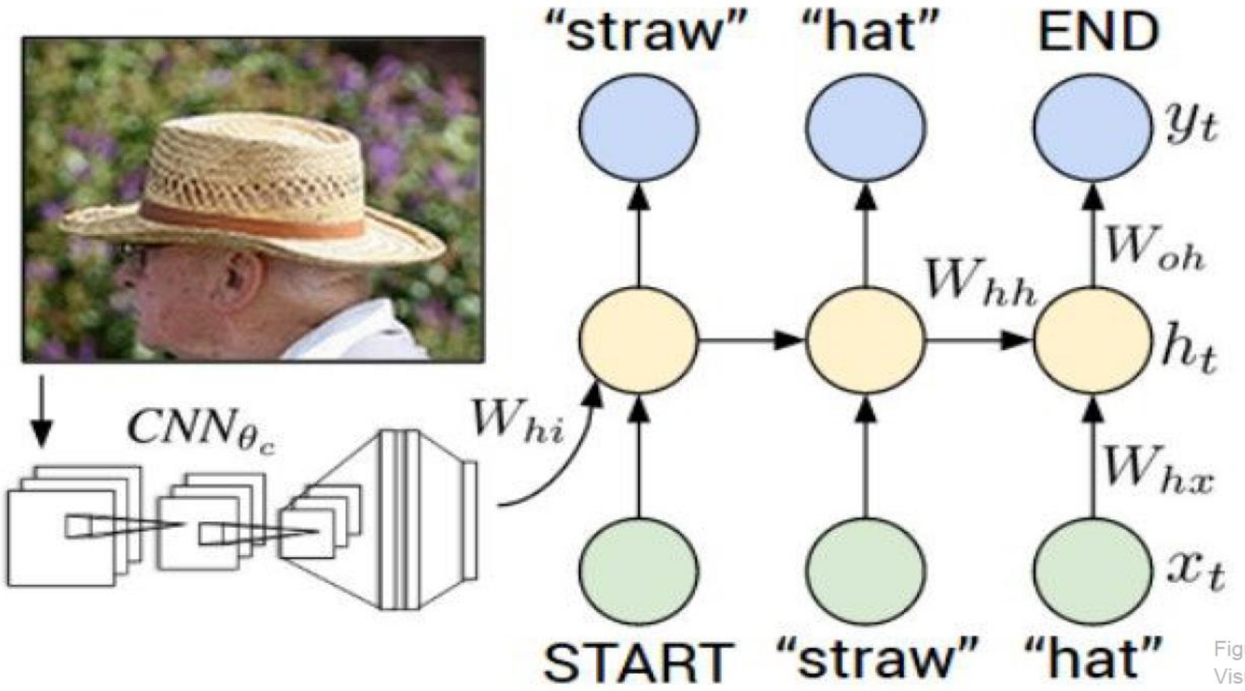
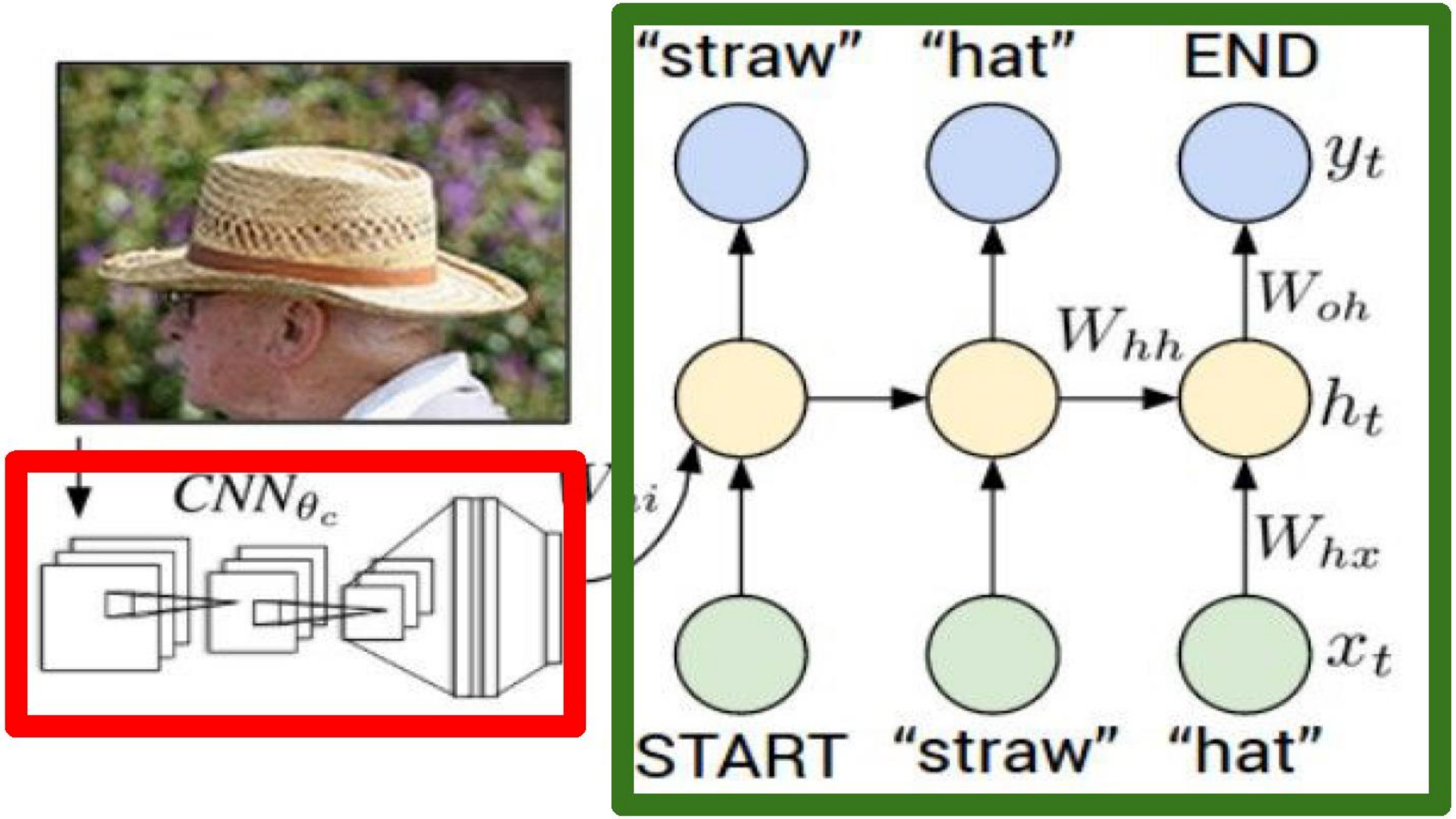


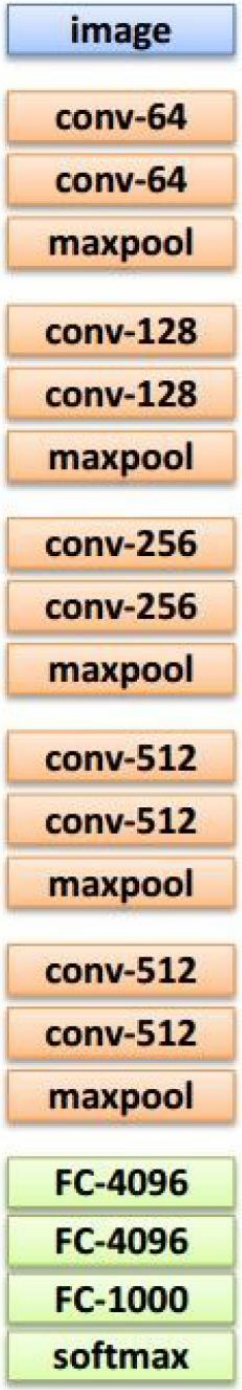
Figure from Karpathy et a, "Deep Visual-Semantic Alignments for Generating Image Descriptions", CVPR 2015; figure copyright IEEE, 2015.
Reproduced for educational purposes.

Explain Images with Multimodal Recurrent Neural Networks, Mao et al.
Deep Visual-Semantic Alignments for Generating Image Descriptions, Karpathy and Fei-Fei
Show and Tell: A Neural Image Caption Generator, Vinyals et al.
Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al.
Learning a Recurrent Visual Representation for Image Caption Generation, Chen and Zitnick

Recurrent Neural Network



Convolutional Neural Network



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

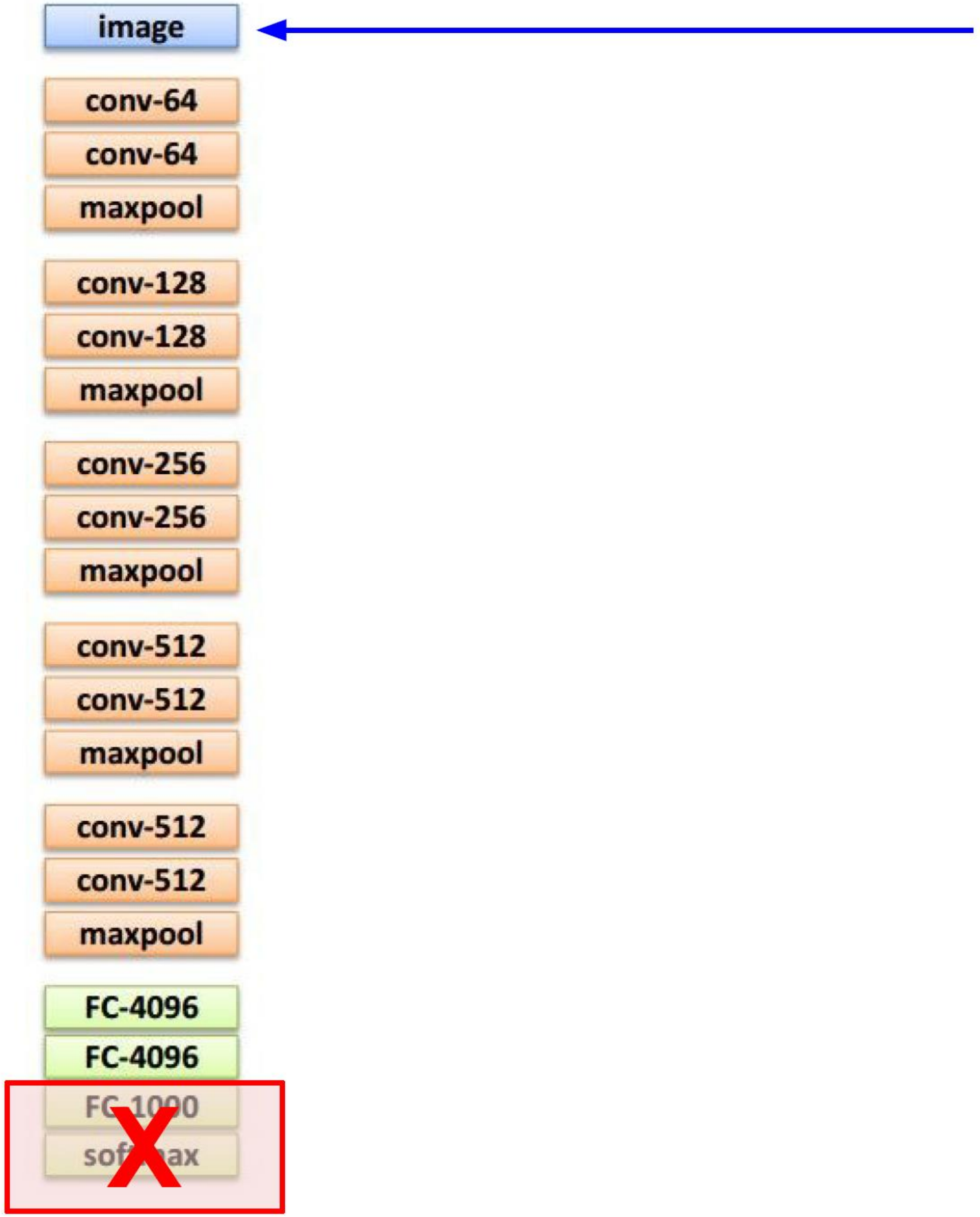
FC-4096

FC-4096

FC-1000

softmax

test image



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

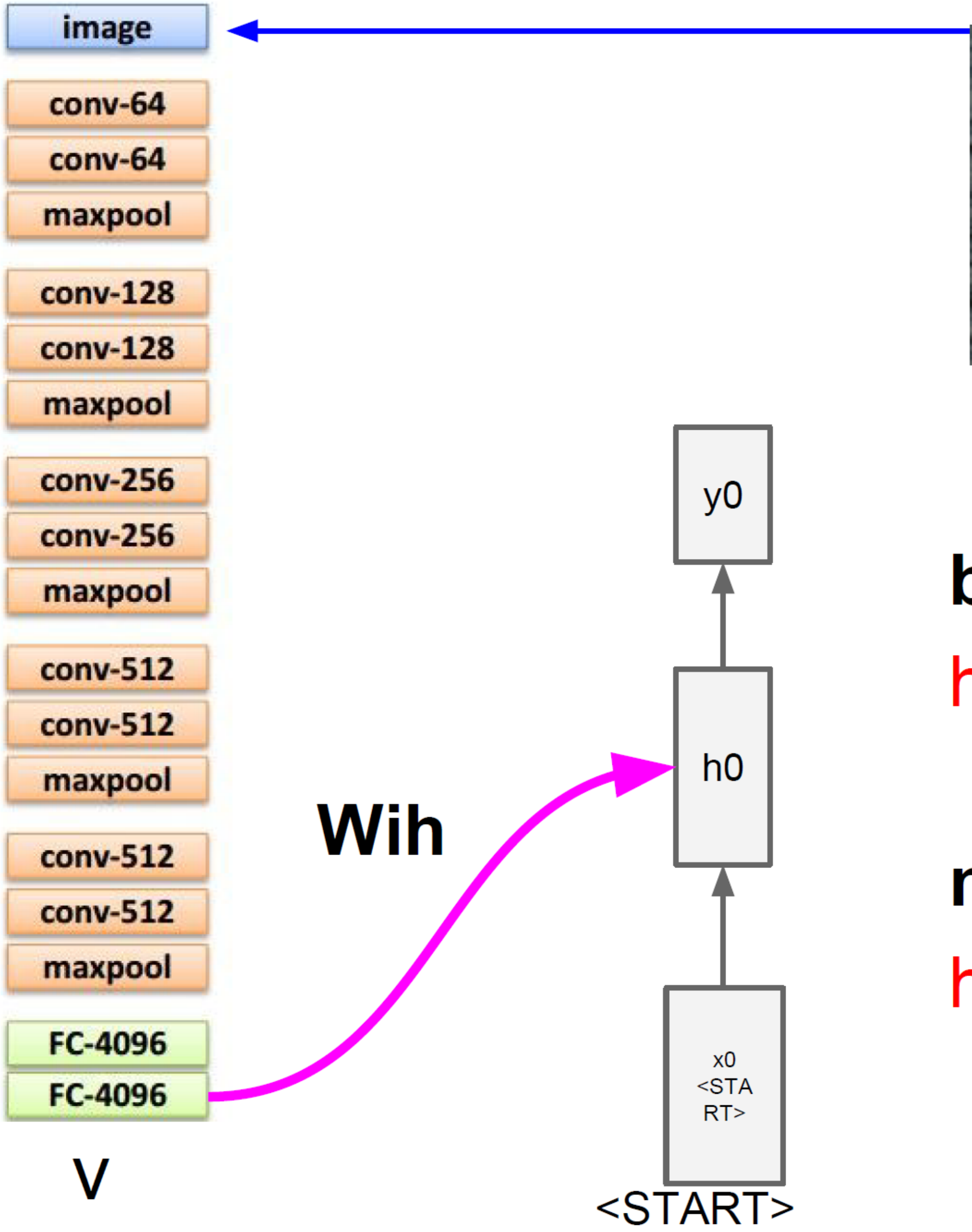
FC-4096

FC-4096

~~FC-1000~~

~~softmax~~

test image



test image

before:

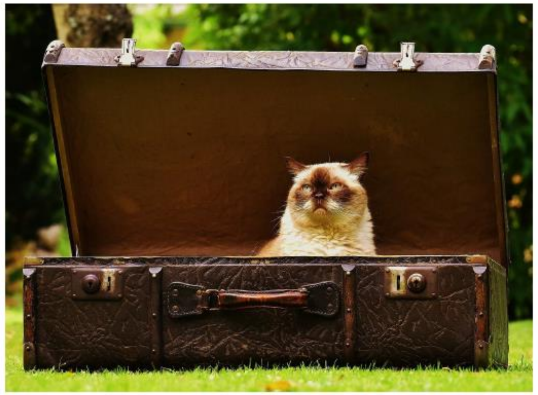
$$h = \tanh(Wxh * x + Whh * h)$$

now:

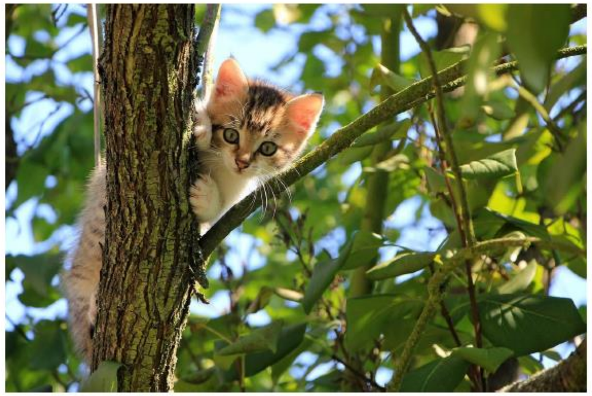
$$h = \tanh(Wxh * x + Whh * h + Wih * v)$$

Image Captioning: Example Results

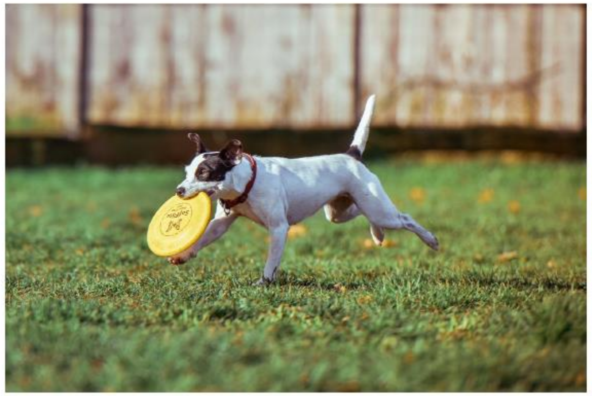
Captions generated using neuraltalk2
All images are CC0 Public domain:
cat suitcase, cat tree, dog, bear,
surfers, tennis, giraffe, motorcycle



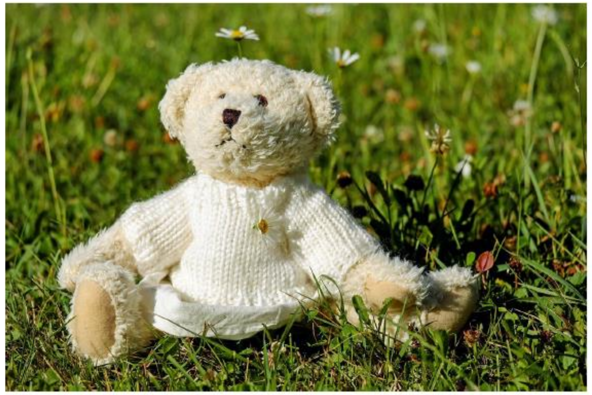
A cat sitting on a suitcase on the floor



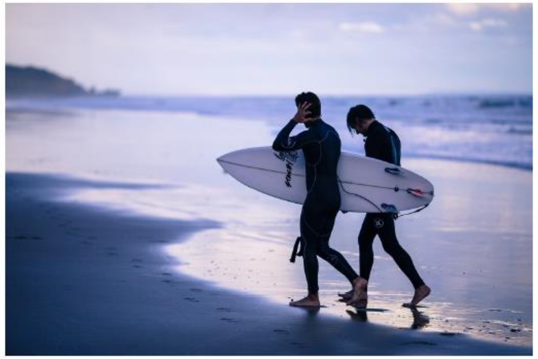
A cat is sitting on a tree branch



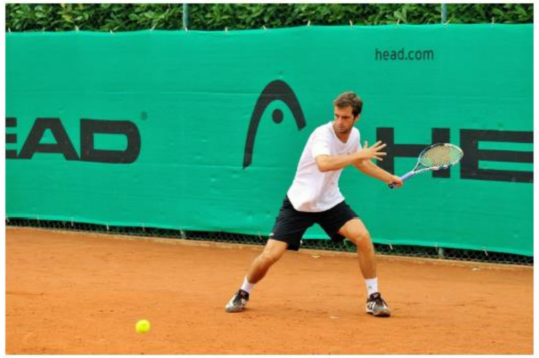
A dog is running in the grass with a frisbee



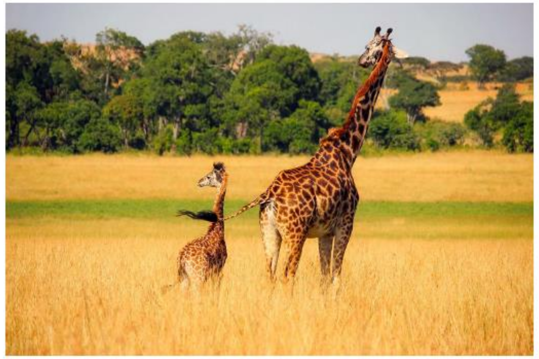
A white teddy bear sitting in the grass



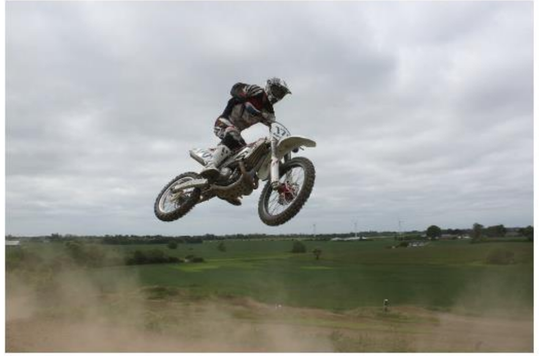
Two people walking on the beach with surfboards



A tennis player in action on the court



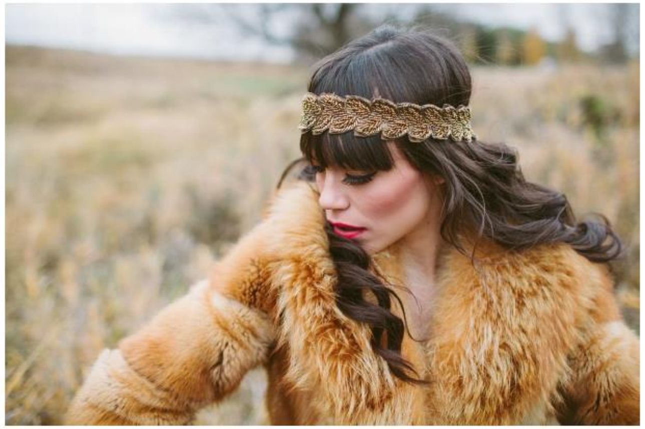
Two giraffes standing in a grassy field



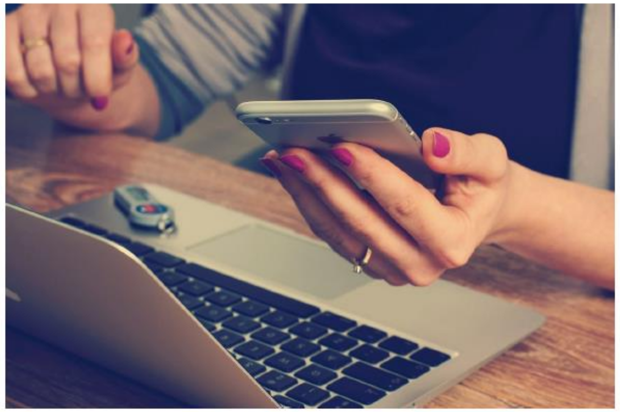
A man riding a dirt bike on a dirt track

Image Captioning: Failure Cases

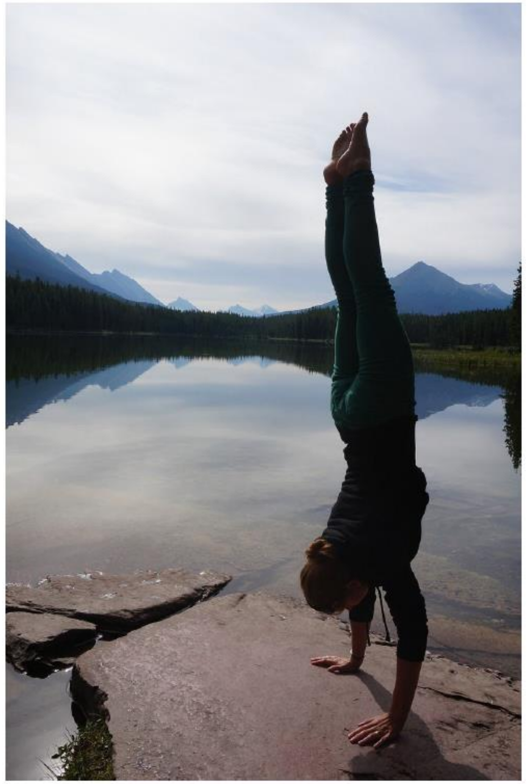
Captions generated using neuraltalk2
All images are CC0 Public domain: fur coat, handstand, spider web, baseball



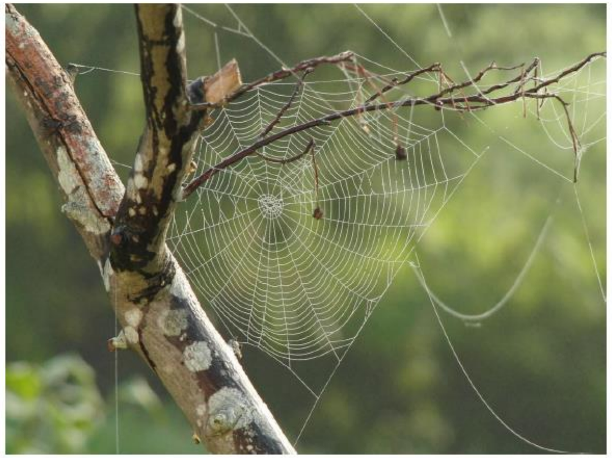
A woman is holding a cat in her hand



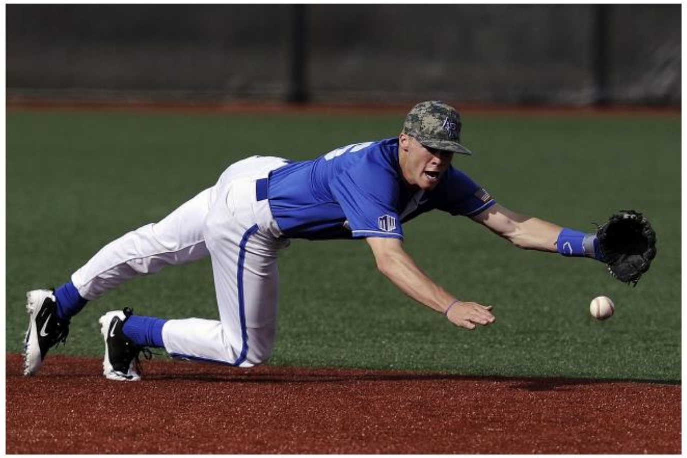
A person holding a computer mouse on a desk



A woman standing on a beach holding a surfboard



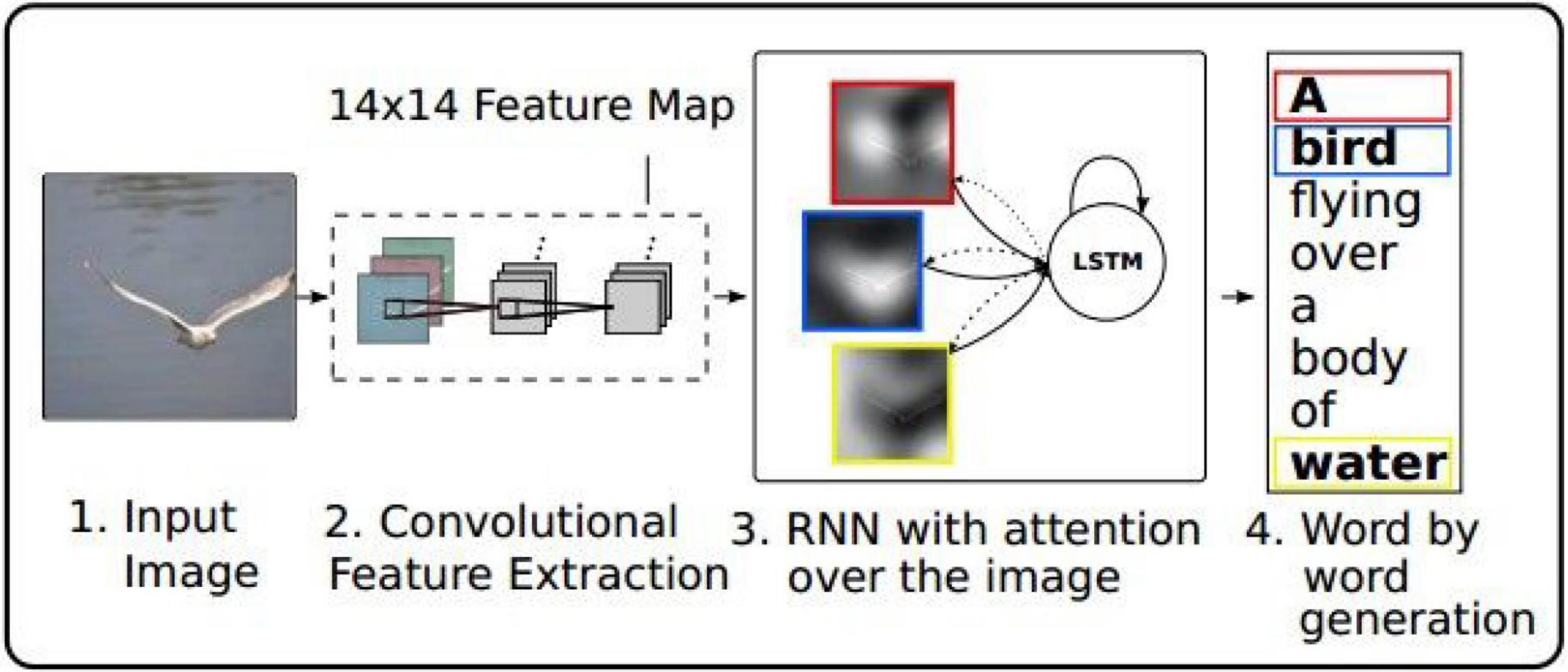
A bird is perched on a tree branch



A man in a baseball uniform throwing a ball

Image Captioning with Attention

RNN focuses its attention at a different spatial location when generating each word


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

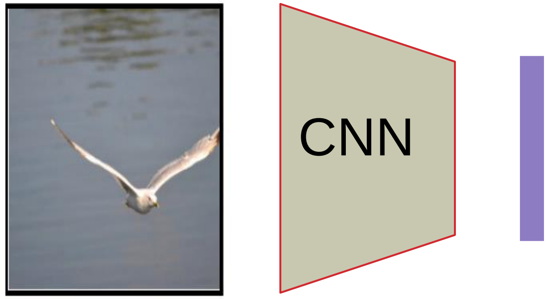
Xu et al, “Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention”, ICML 2015

Figure copyright Kelvin Xu, Jimmy Lei Ba, Jamie Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Benchio, 2015. Reproduced with permission.

Image Captioning

Image:
H x W x 3

Image Captioning

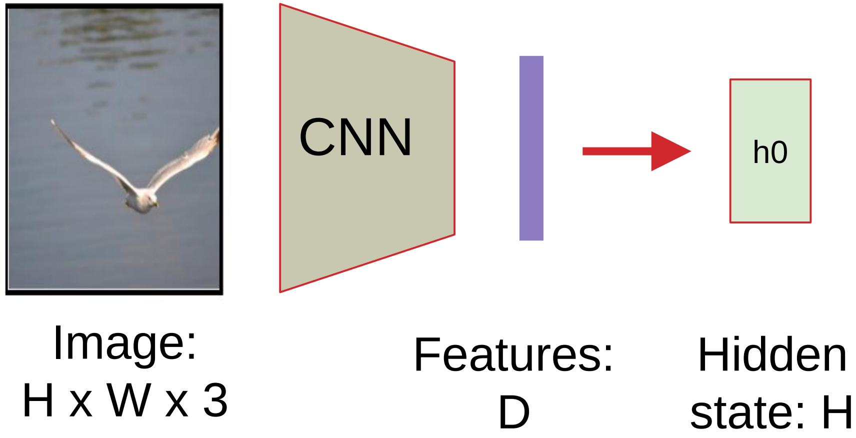


CNN

Image:
H x W x 3

Features:
D

Image Captioning



CNN

h0

Image:
H x W x 3

Features:
D

Hidden
state: H

Image Captioning

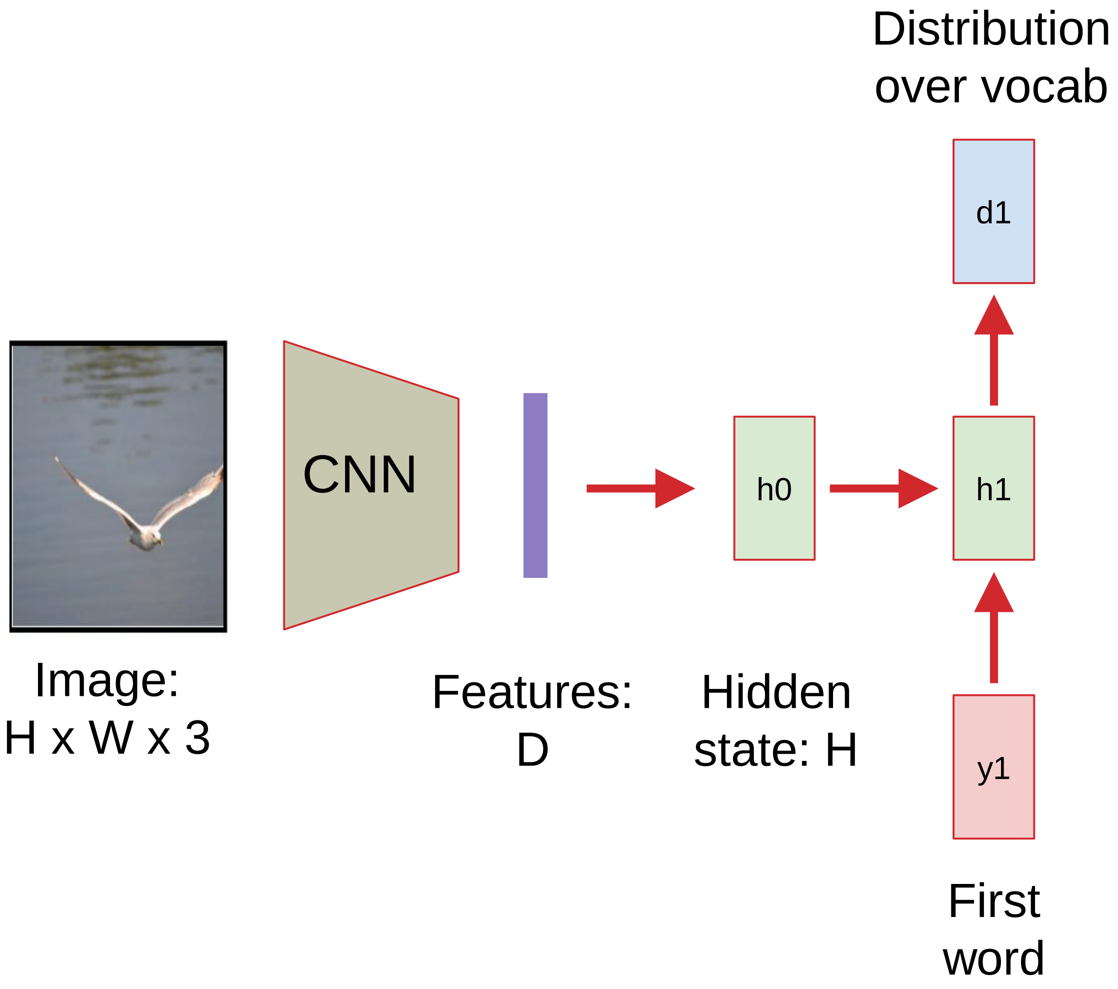
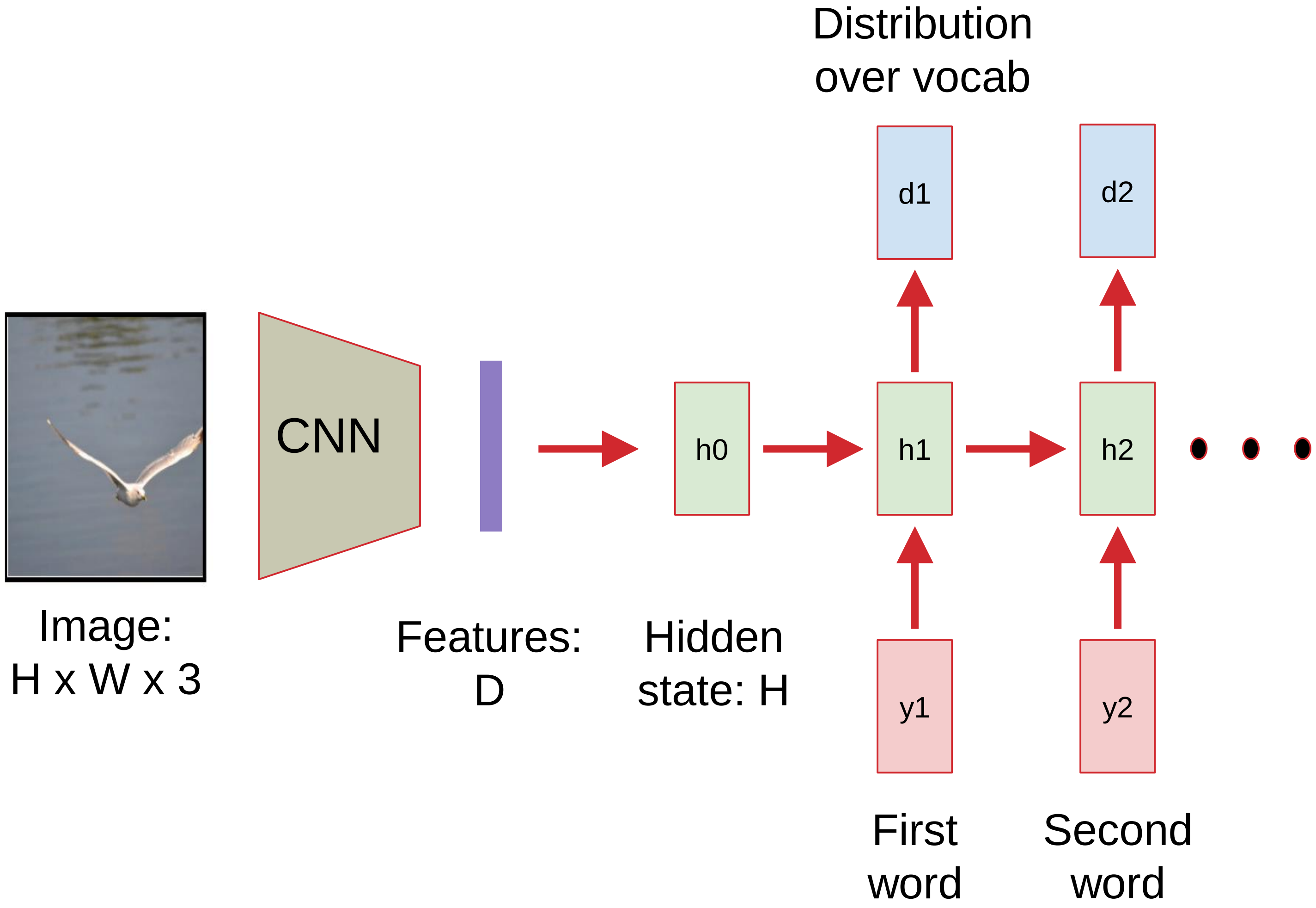


Image Captioning



Distribution over vocab

Image:
H x W x 3

Features:
D

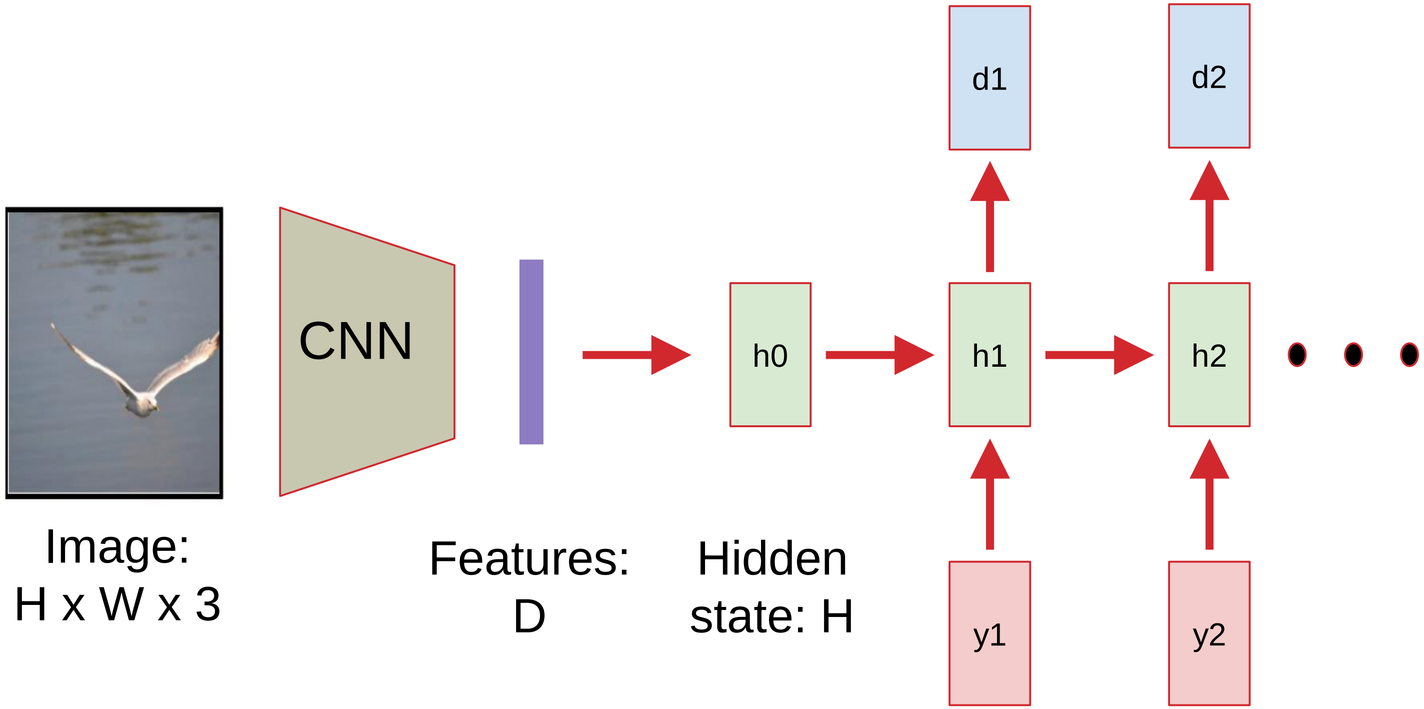
Hidden
state: H

First
word

Second
word

Image Captioning

Distribution over vocab



d1

d2

CNN

h0

h1

h2

y1

y2

Image: H x W x 3

Features: D

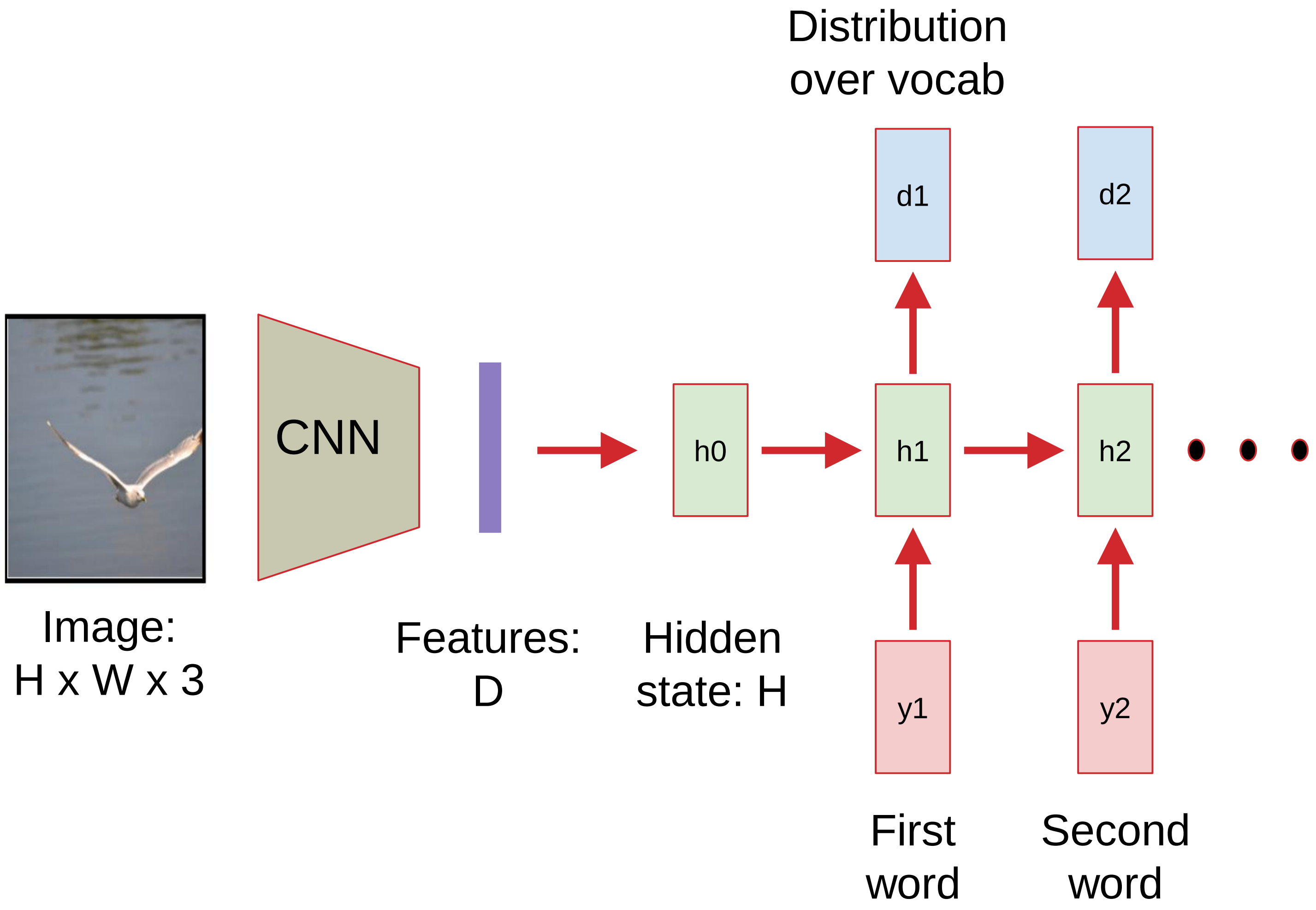
Hidden state: H

First word

Second word

RNN only looks at whole image, once

Image Captioning



Distribution over vocab

d1

d2

CNN

h0

h1

h2

y1

y2

Image: H x W x 3

Features: D

Hidden state: H

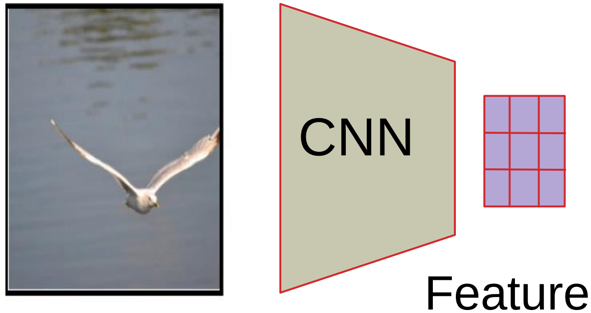
First word

Second word

RNN only looks at whole image, once

What if the RNN looks at different parts of the image at each timestep?

Image Captioning with Attention



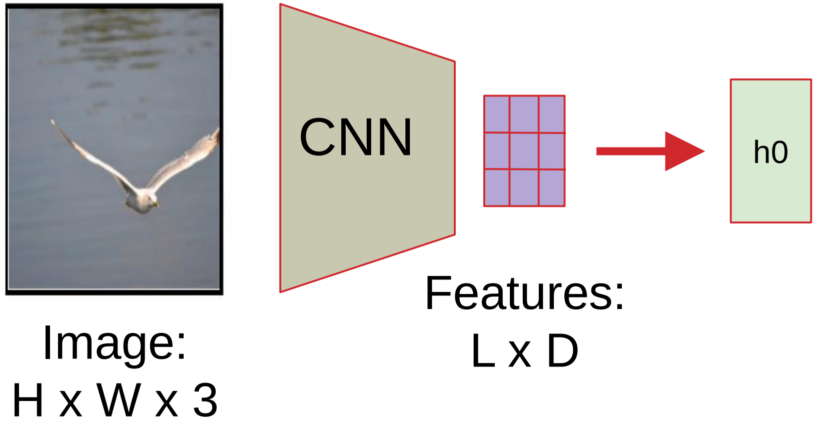
CNN

Features:
L x D

Image:
H x W x 3

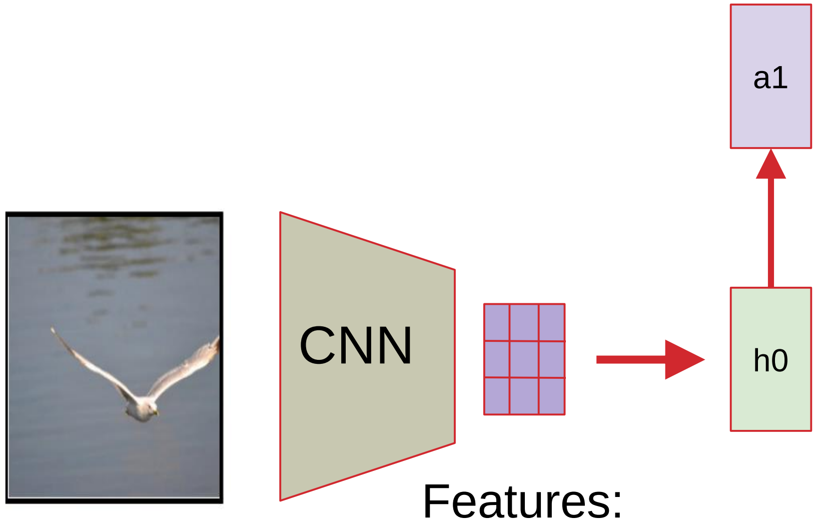
Xu et al, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML 2015

Image Captioning with Attention



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

Image Captioning with Attention



Distribution over
L locations

a1

CNN

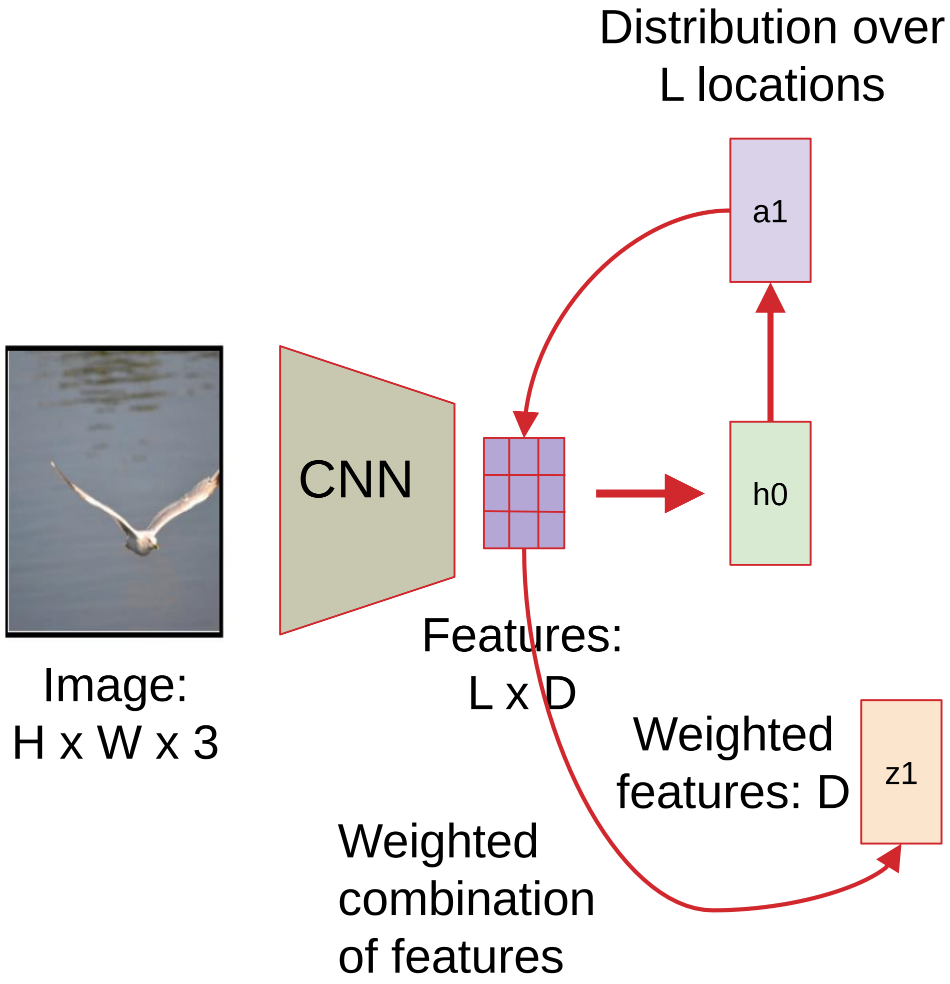
h0

Image:
H x W x 3

Features:
L x D

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

Image Captioning with Attention



Distribution over L locations

Image: H x W x 3

Features: L x D

Weighted features: D

Weighted combination of features

Image Captioning with Attention

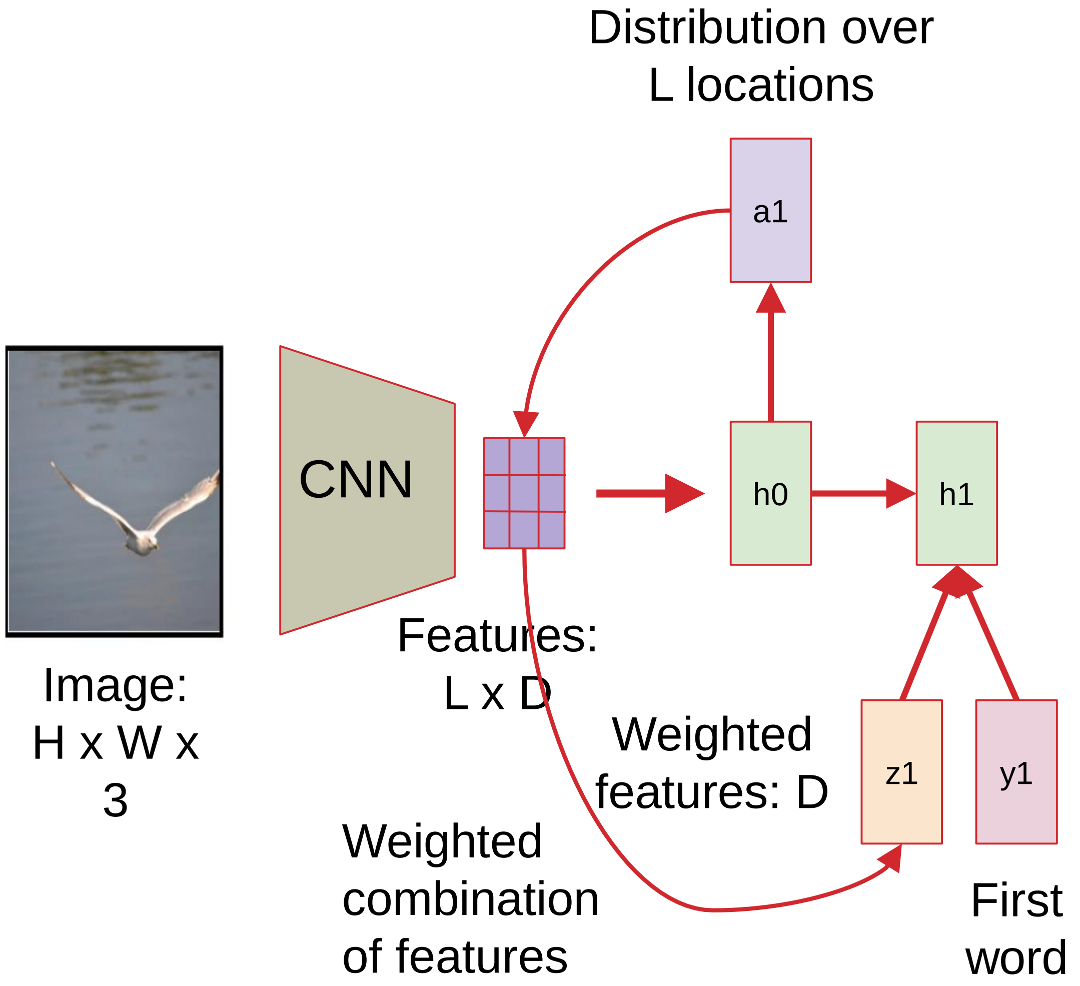


Image Captioning with Attention

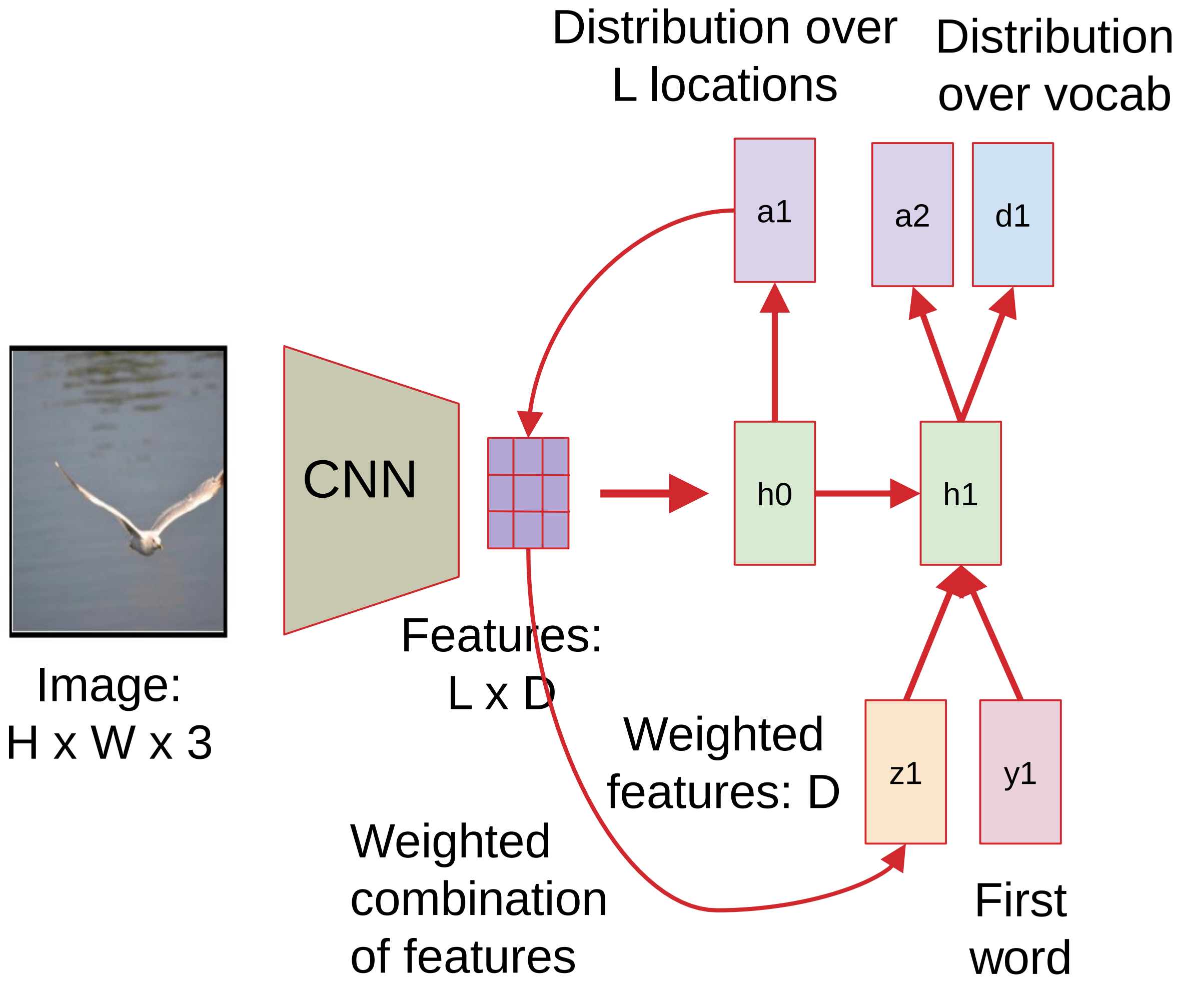


Image Captioning with Attention

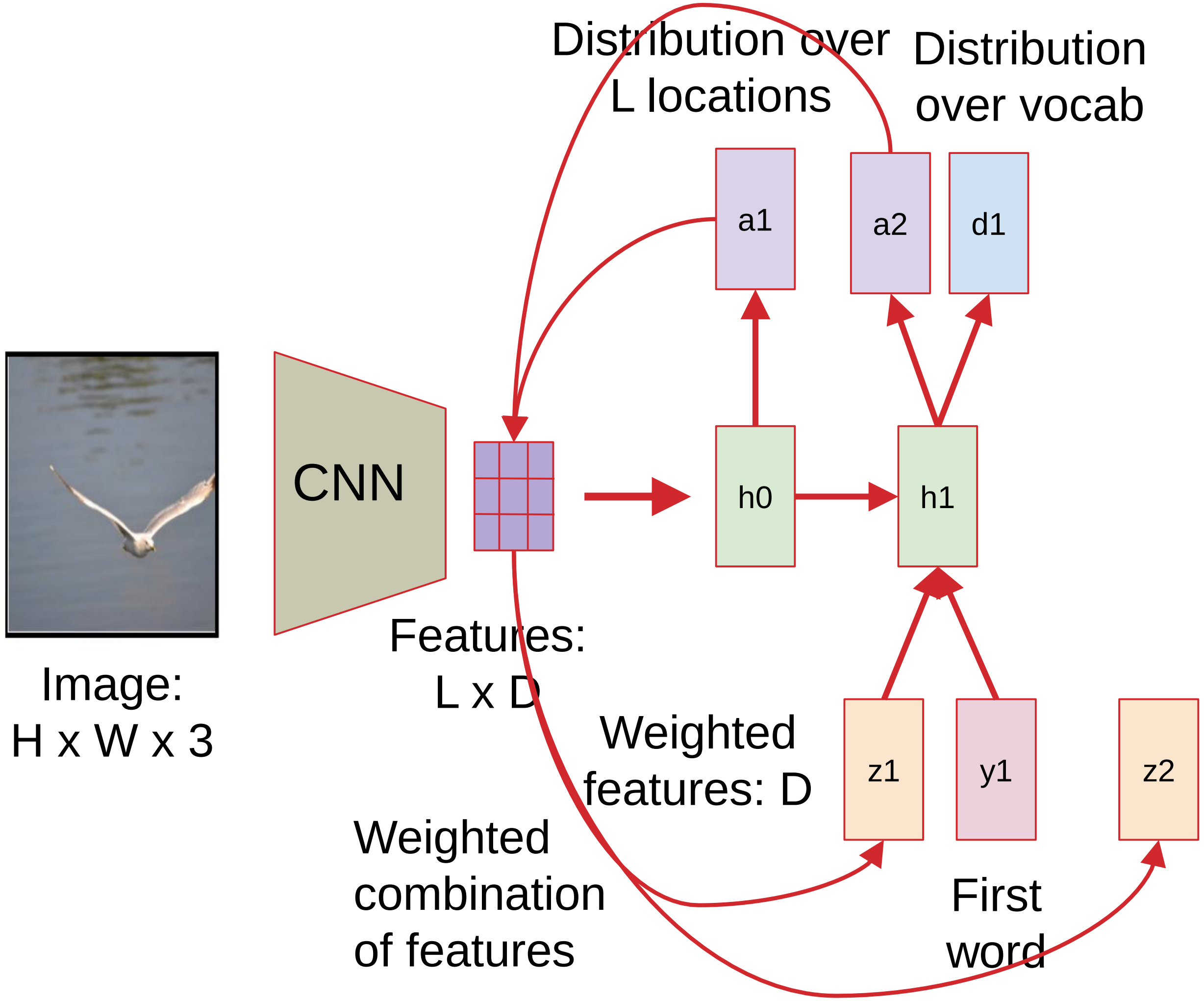


Image Captioning with Attention

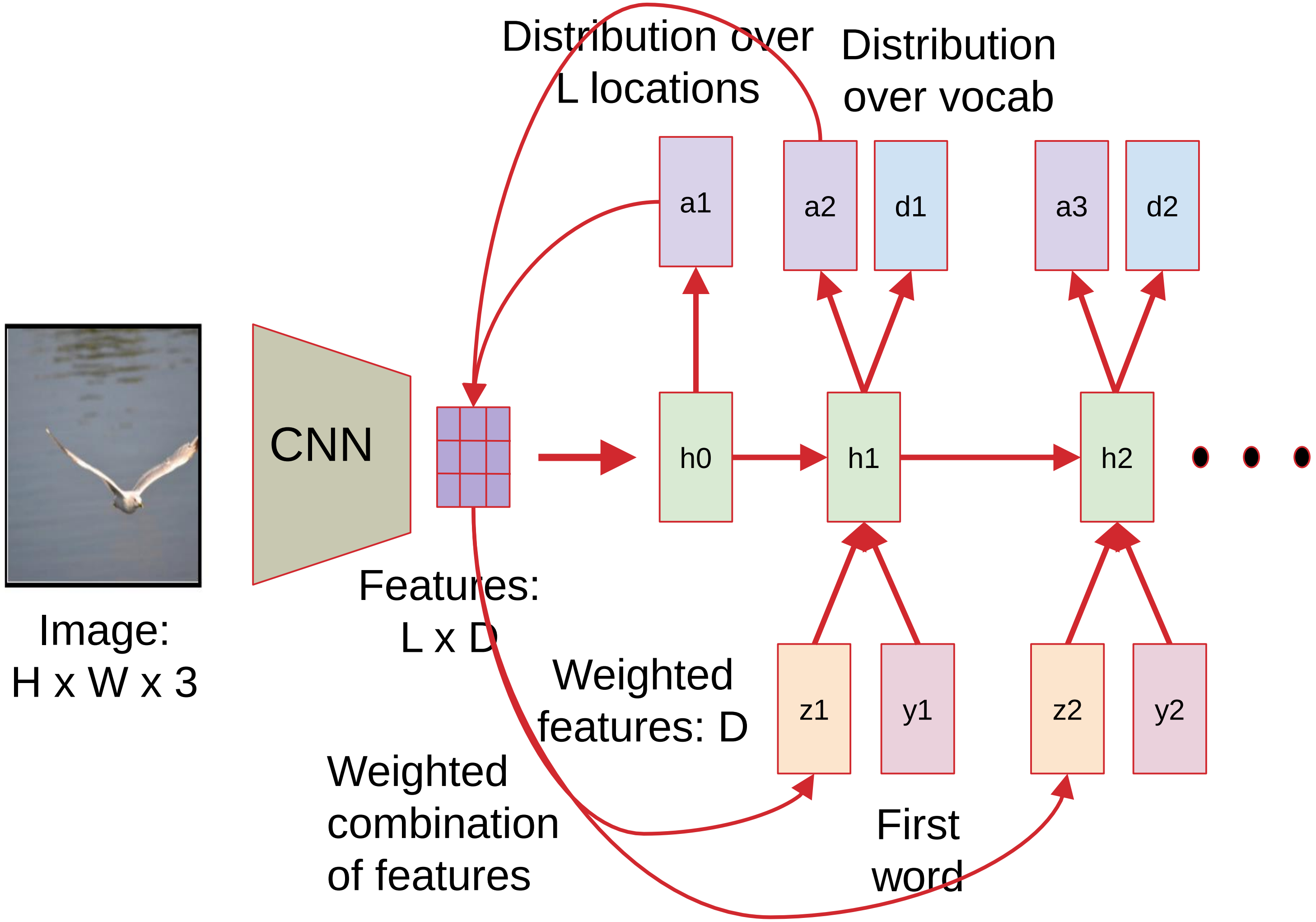


Image Captioning with Attention

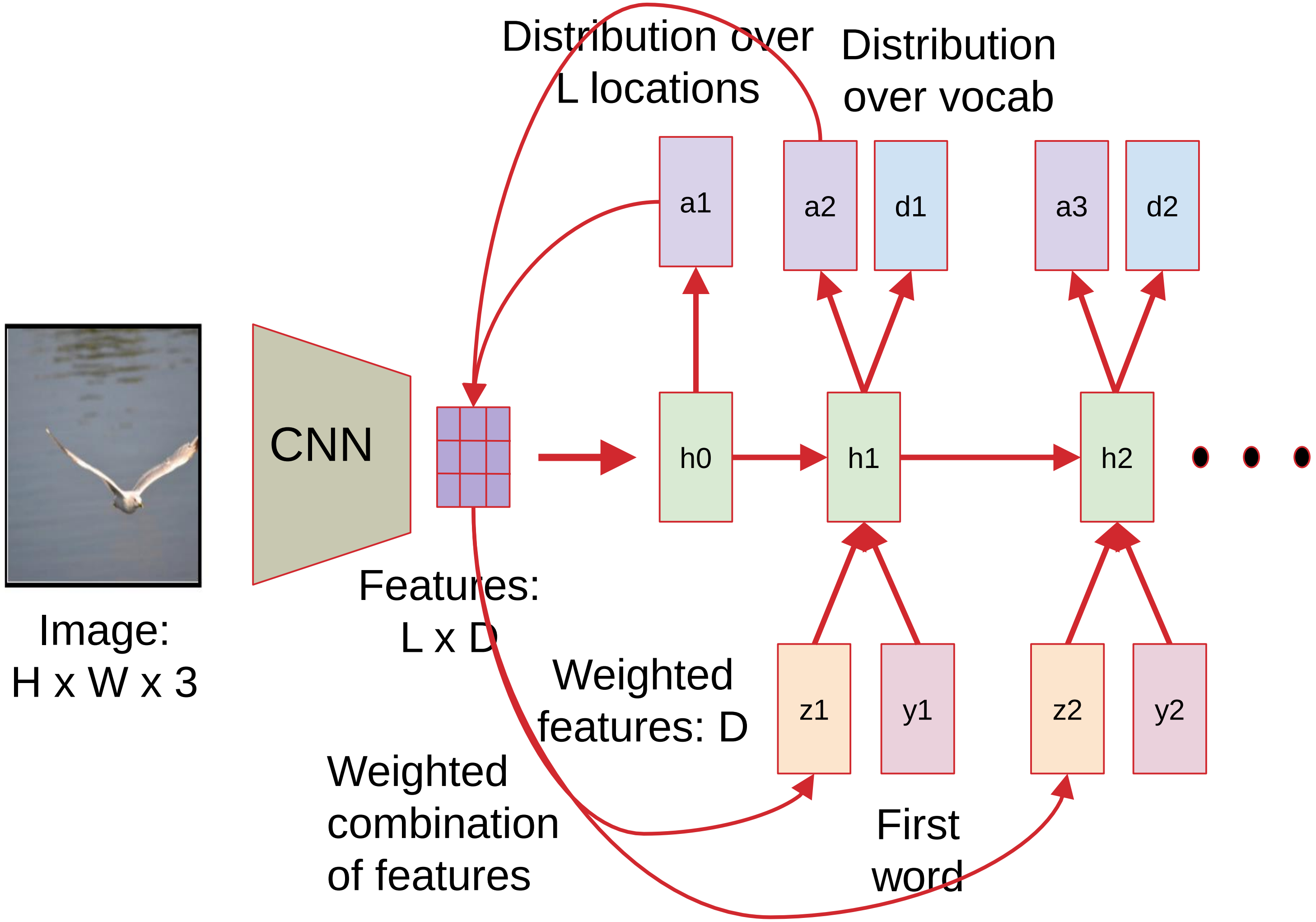


Image Captioning with Attention

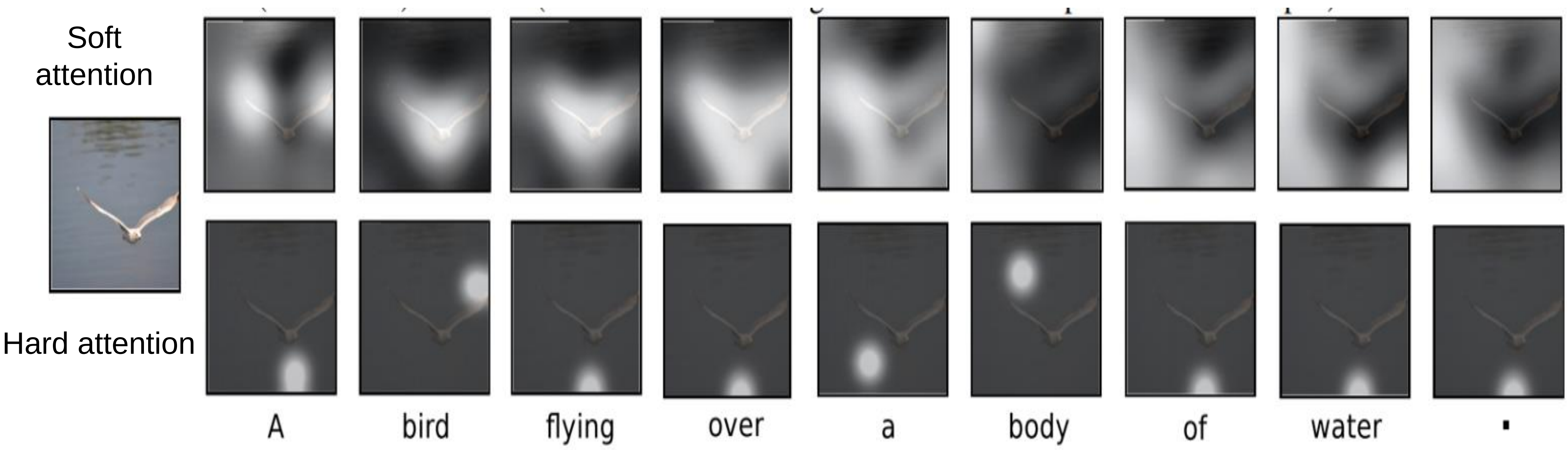
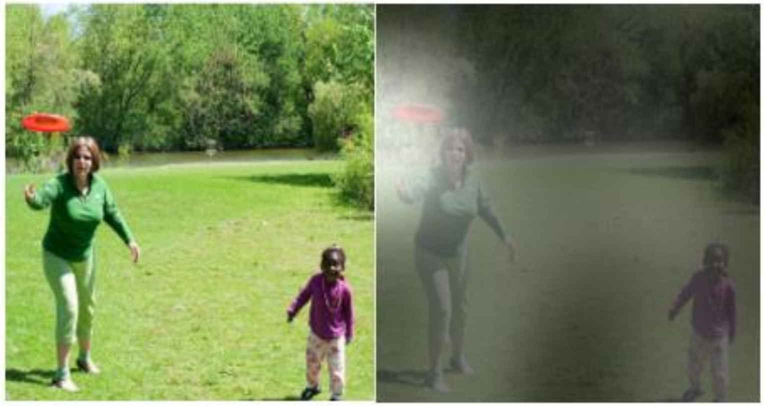
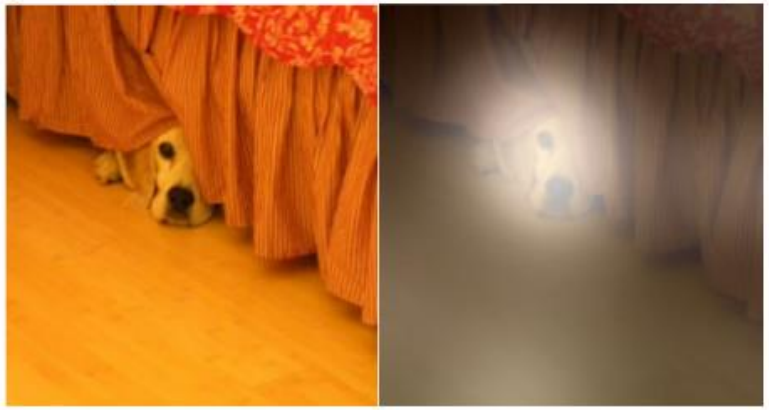


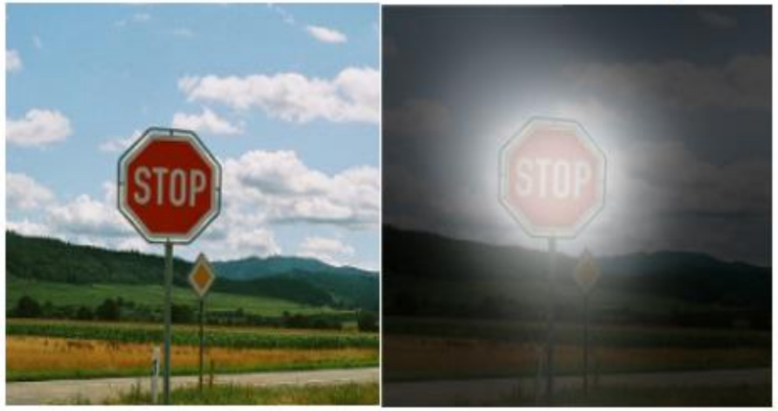
Image Captioning with Attention



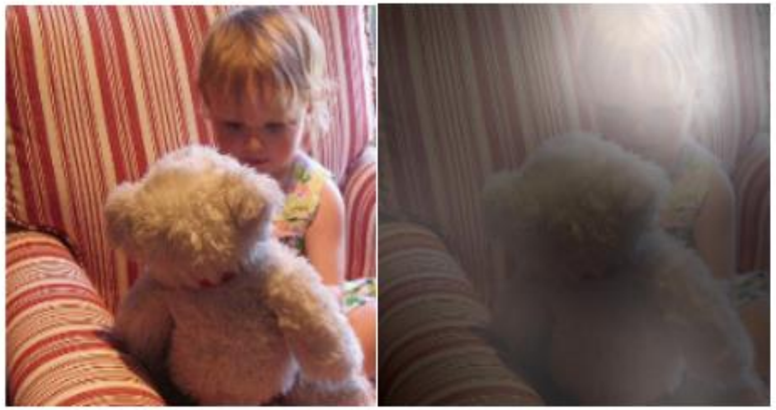
A woman is throwing a frisbee in a park.



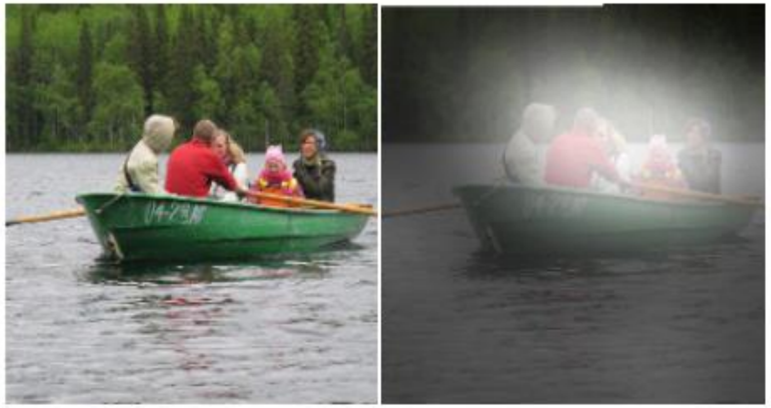
A dog is standing on a hardwood floor.



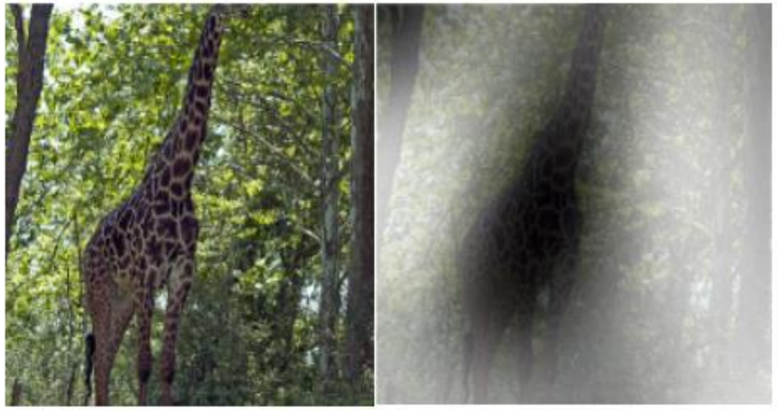
A stop sign is on a road with a mountain in the background.



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

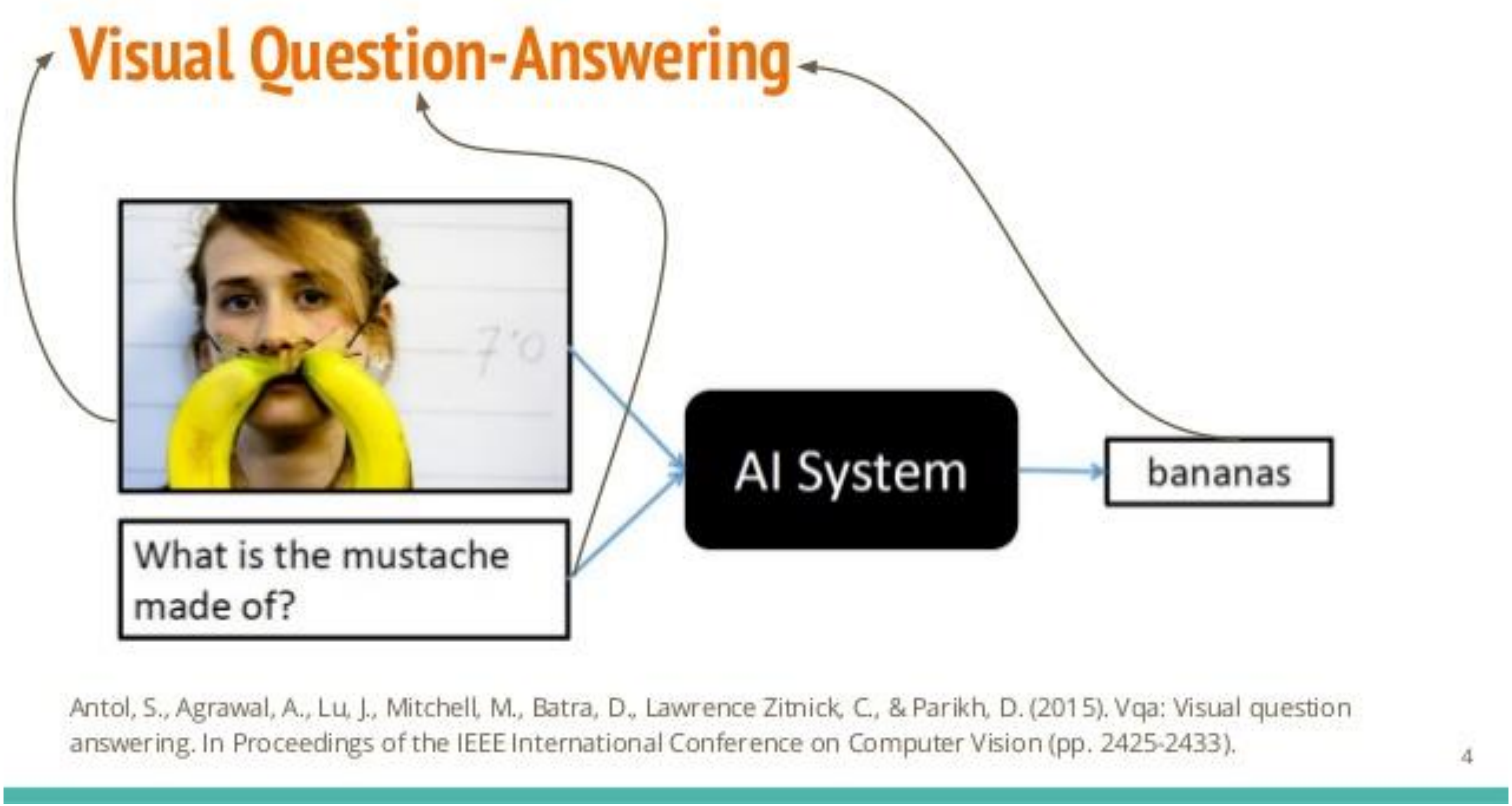


A group of people sitting on a boat in the water.



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

Visual Question Answering



Visual Question-Answering

What is the mustache made of?

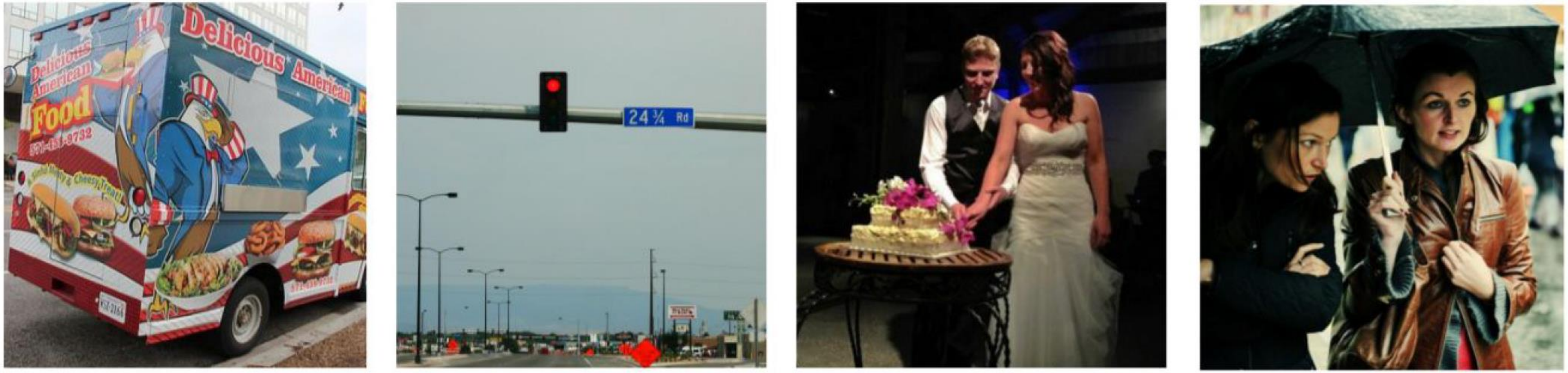
AI System

bananas

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2425-2433).

4

Visual Question Answering



Q: What endangered animal is featured on the truck?

A: **A bald eagle.**
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?

A: **Onto 24 ¾ Rd.**
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?

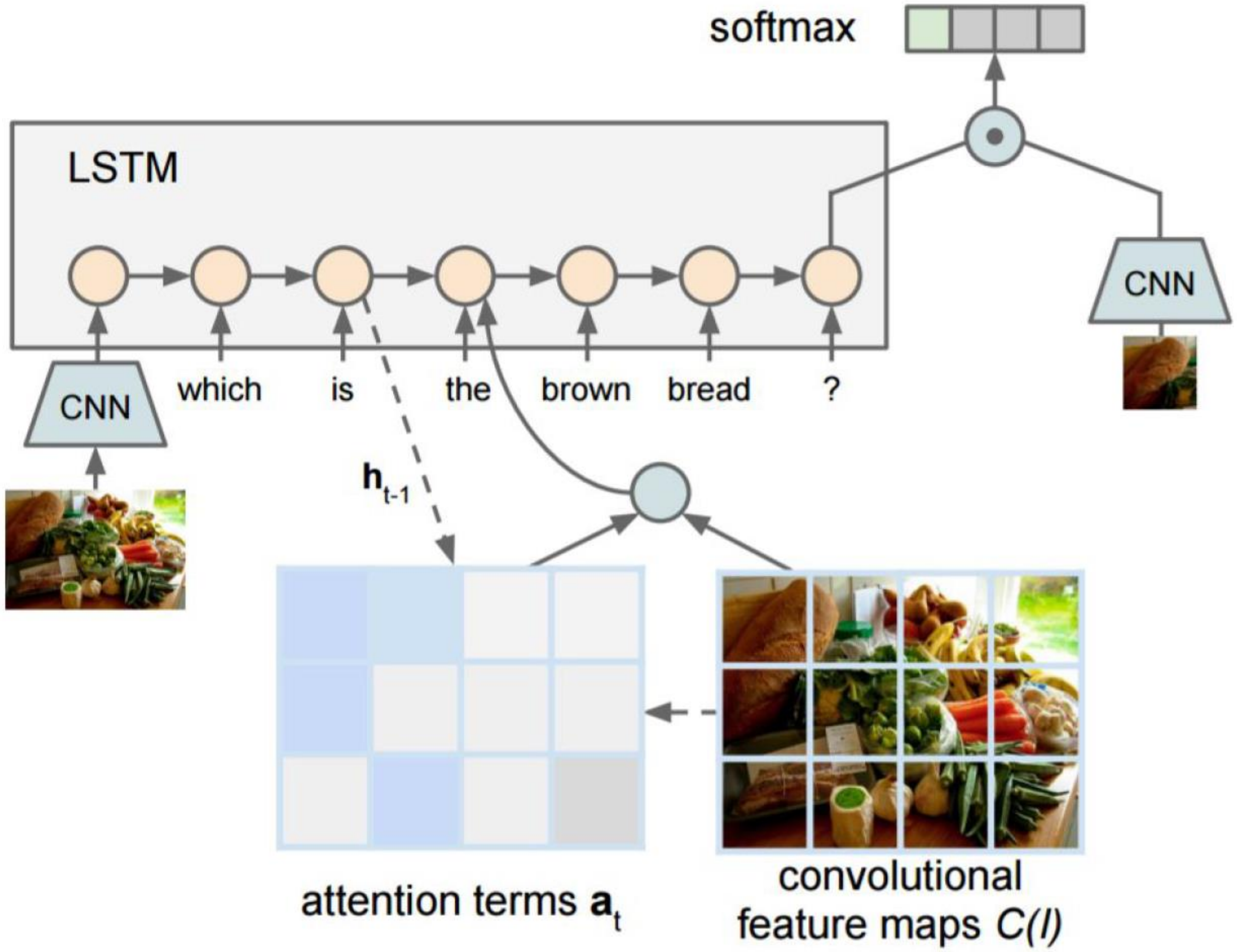
A: **During a wedding.**
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service

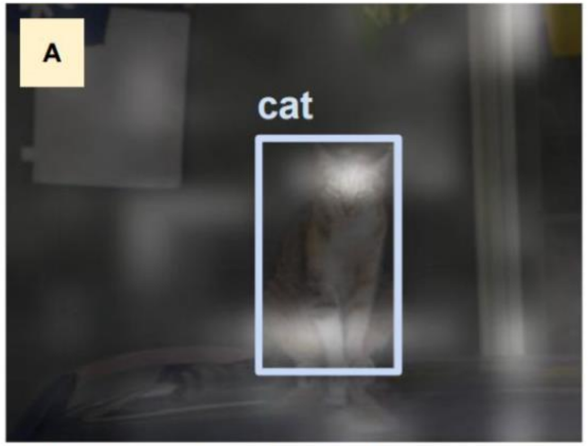
Q: Who is under the umbrella?

A: **Two women.**
A: A child.
A: An old man.
A: A husband and a wife.

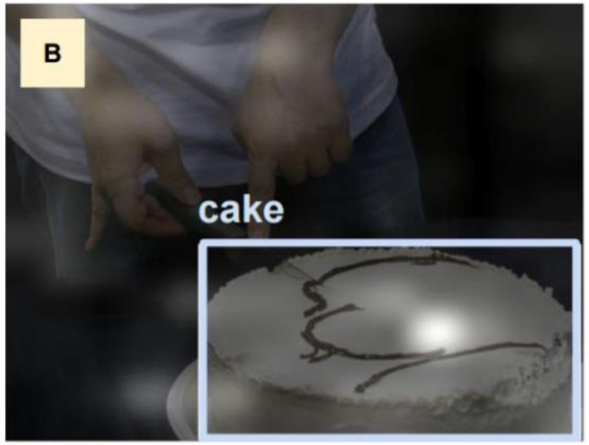
Agrawal et al, "VQA: Visual Question Answering", ICCV 2015
Zhu et al, "Visual 7W: Grounded Question Answering in Images", CVPR 2016
Figure from Zhu et al, copyright IEEE 2016. Reproduced for educational purposes.

Visual Question Answering: RNNs with Attention


softmax
LSTM
CNN
which
is
the
brown
bread
?
CNN
h_{t-1}
attention terms $\mathbf{a}_t$
convolutional feature maps $C(I)$



What kind of animal is in the photo?
A **cat**.



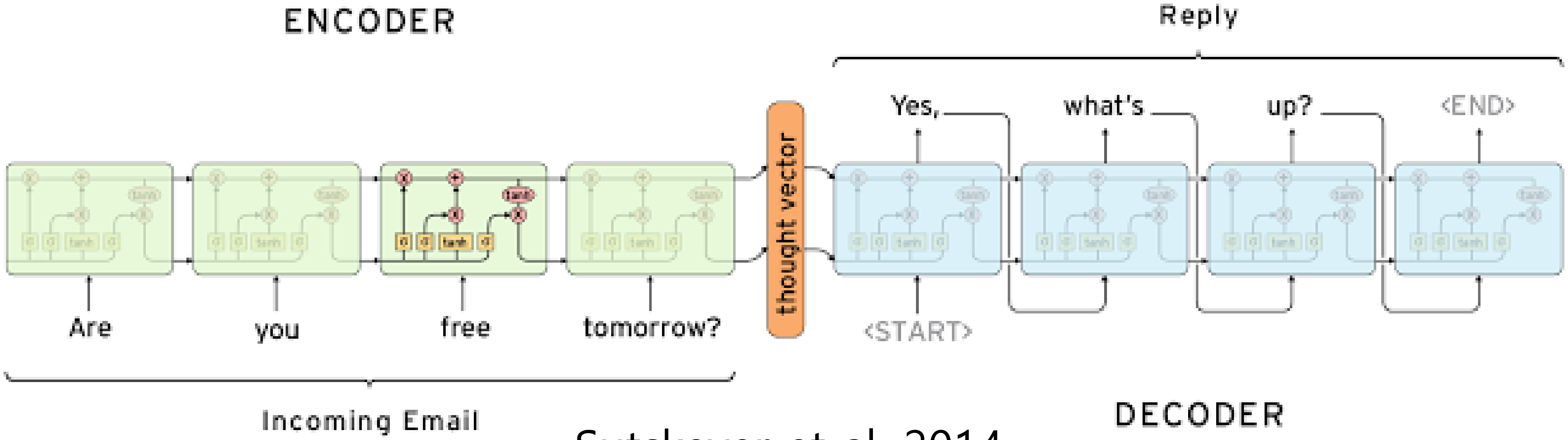
Why is the person holding a knife?
To cut the **cake** with.

Zhu et al, "Visual 7W: Grounded Question Answering in Images", CVPR 2016

Figures from Zhu et al, copyright IEEE 2016. Reproduced for educational purposes.

Sequence to Sequence Model (seq2seq)

- 시퀀스를 입력으로 받아서, 시퀀스를 출력으로 생성
- 많은 NLP task 들에서 기본 모델로 활용됨: 챗봇, 기계번역 등



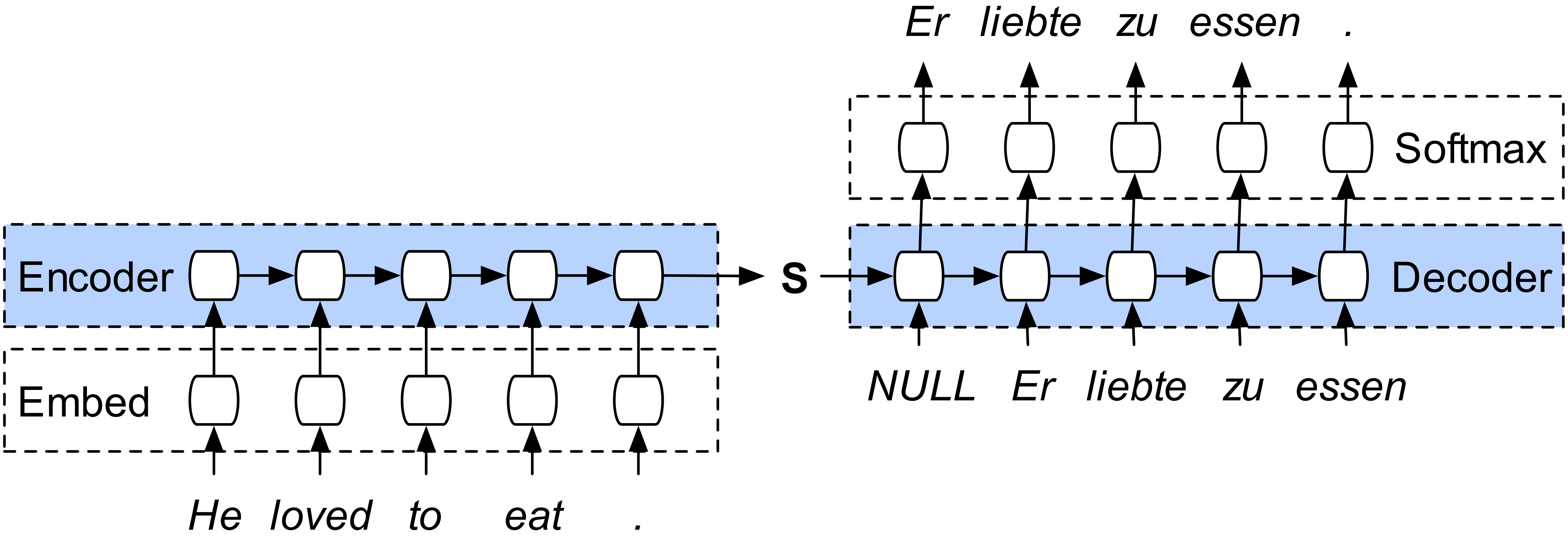
Sutskever et al. 2014

"Sequence to Sequence Learning with Neural Networks"

Encode source into fixed length vector, use it as initial recurrent state for target decoder model

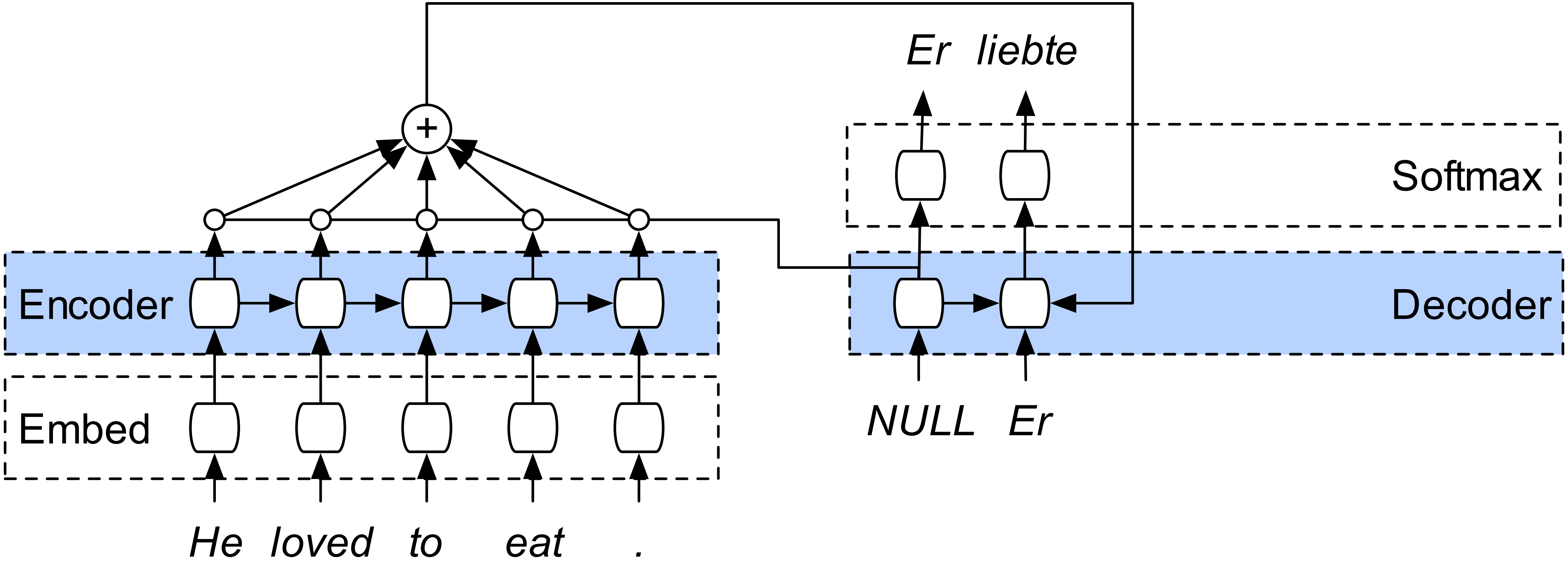
Seq2seq for Machine Translation

- 학습 데이터에서 입력 시퀀스-출력 시퀀스를 번역 데이터로 사용



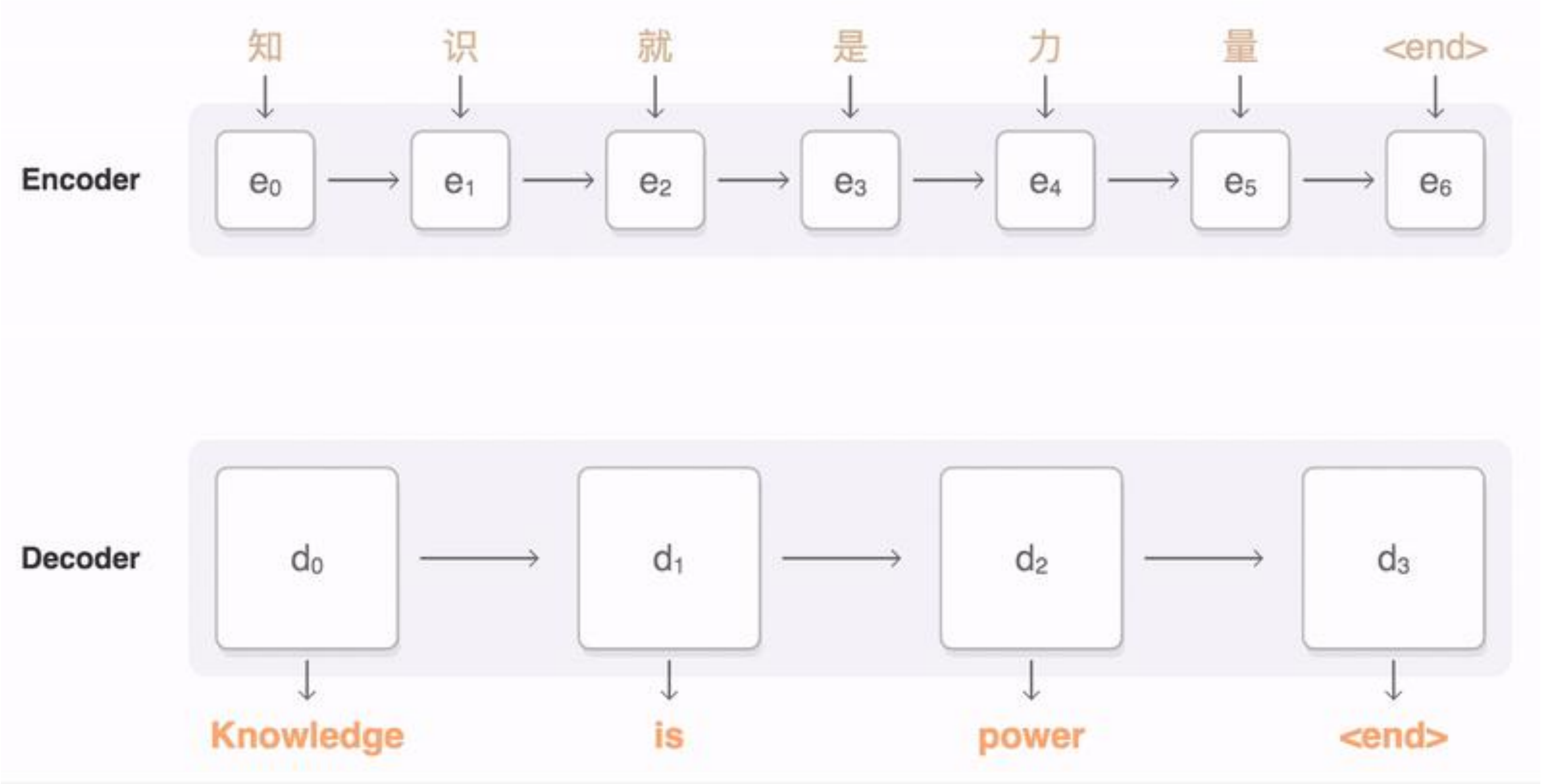
Seq2seq with Attention

- 입력 시퀀스의 마지막 시점의 벡터에 모든 정보를 다 담기가 버거우므로, 모든 입력 시퀀스의 정보를 조합하여 각 출력 단어를 생성
- 기계 번역 예



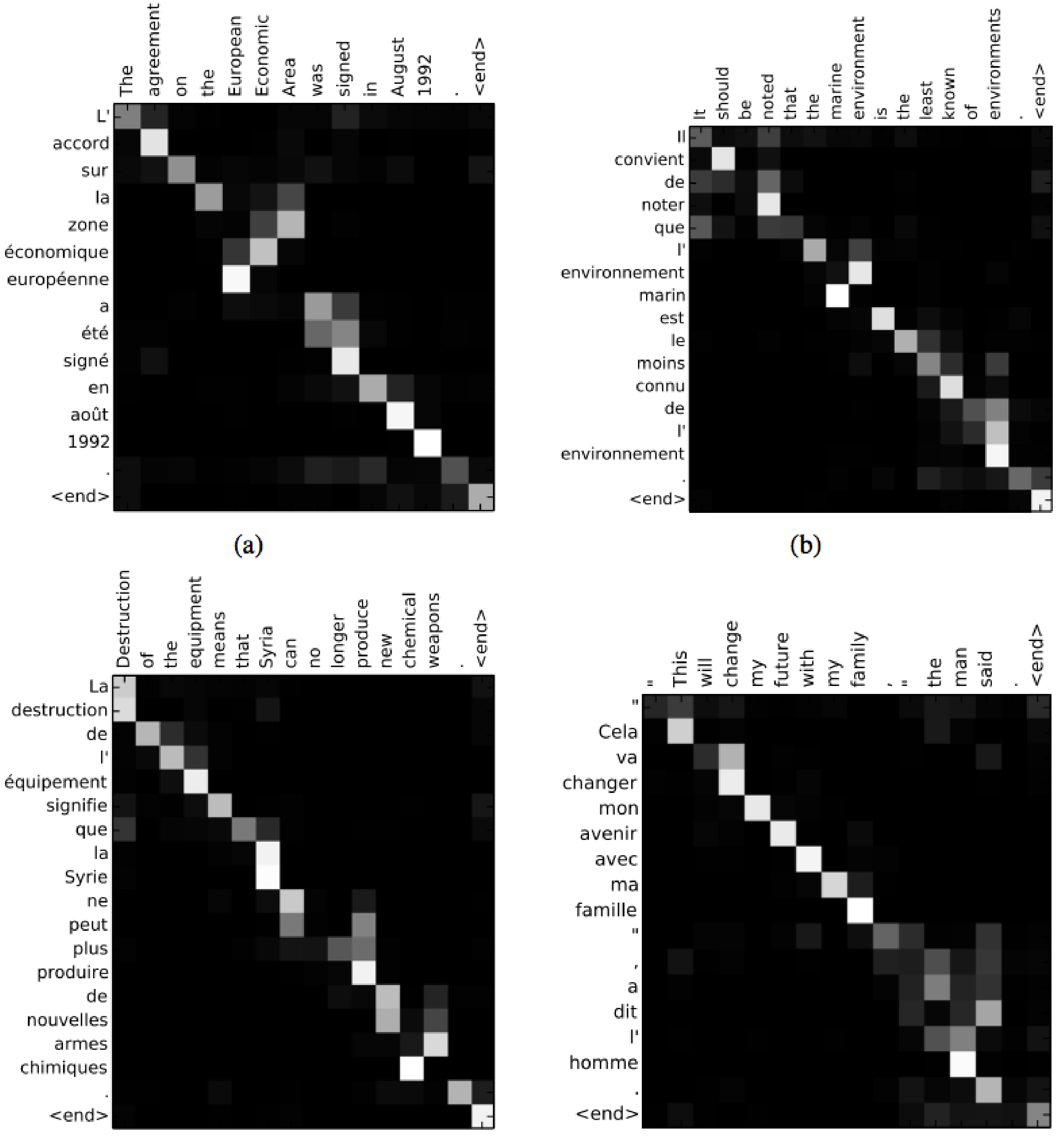
Seq2seq with Attention

- 입력 시퀀스의 마지막 시점의 벡터에 모든 정보를 다 담기가 버거우므로, 모든 입력 시퀀스의 정보를 조합하여 각 출력 단어를 생성
- 기계 번역 예: https://github.com/google/seq2seq



Attention Example in Machine Translation

- 다른 언어들 간의 어순을 학습함
- 관사 등의 필요없는 단어는 건너뜀



(a) (b) (c) (d)

References

Stanford University CS231n: Convolutional Neural Networks for Visual Recognition

Deep Learning Summer School, Montreal 2016 - VideoLectures.NET

Understanding LSTM Networks -- colah's blog