Caption generation

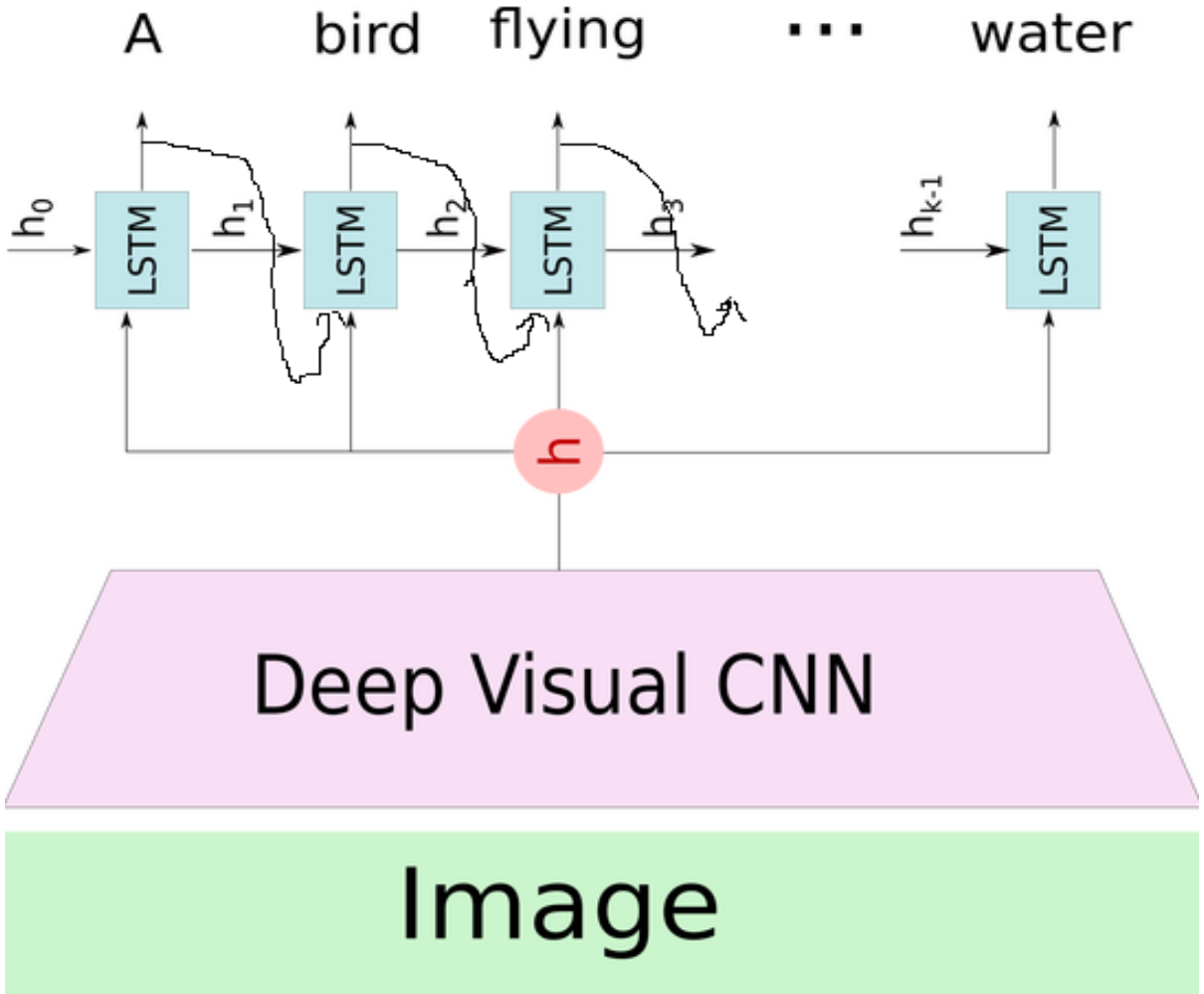
What is an image caption generation?

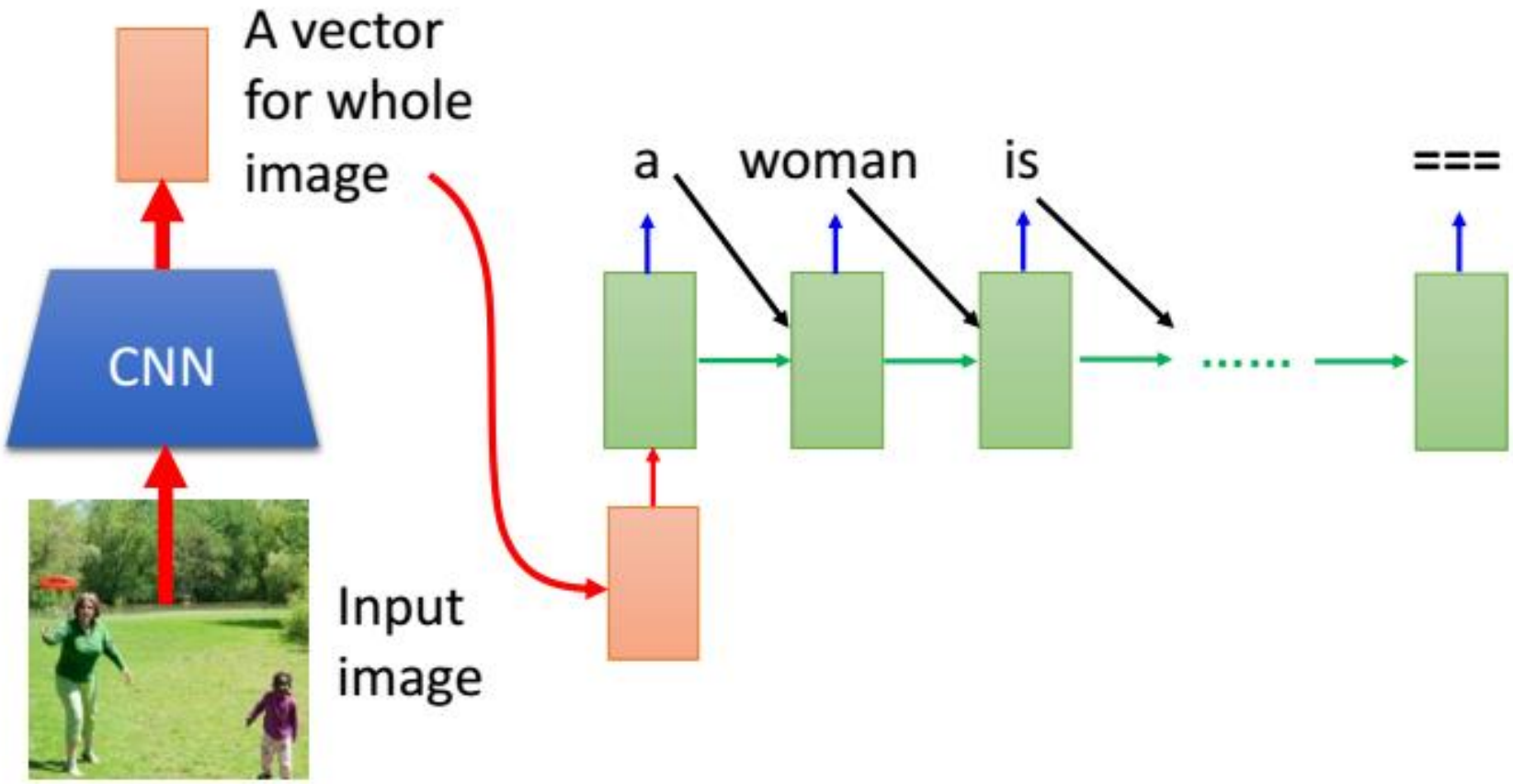
- The task we want to achieve is image captioning: we want to generate a caption for a given image.

Classic Solution

- A « classic » image captioning system would encode the image, using a pre-trained Convolutional Neural Network that would produce a hidden state.

- Then, it would decode this hidden state by using a Recurrent Neural Network (RNN), and generate recursively each word of the caption.


A
bird
flying
...
water
h_0
LSTM
h_1
LSTM
h_2
LSTM
h_3
h_{k-1}
LSTM
h
Deep Visual CNN
Image


A vector for whole image
CNN
Input image
a
woman
is
===

Issues with classic solution

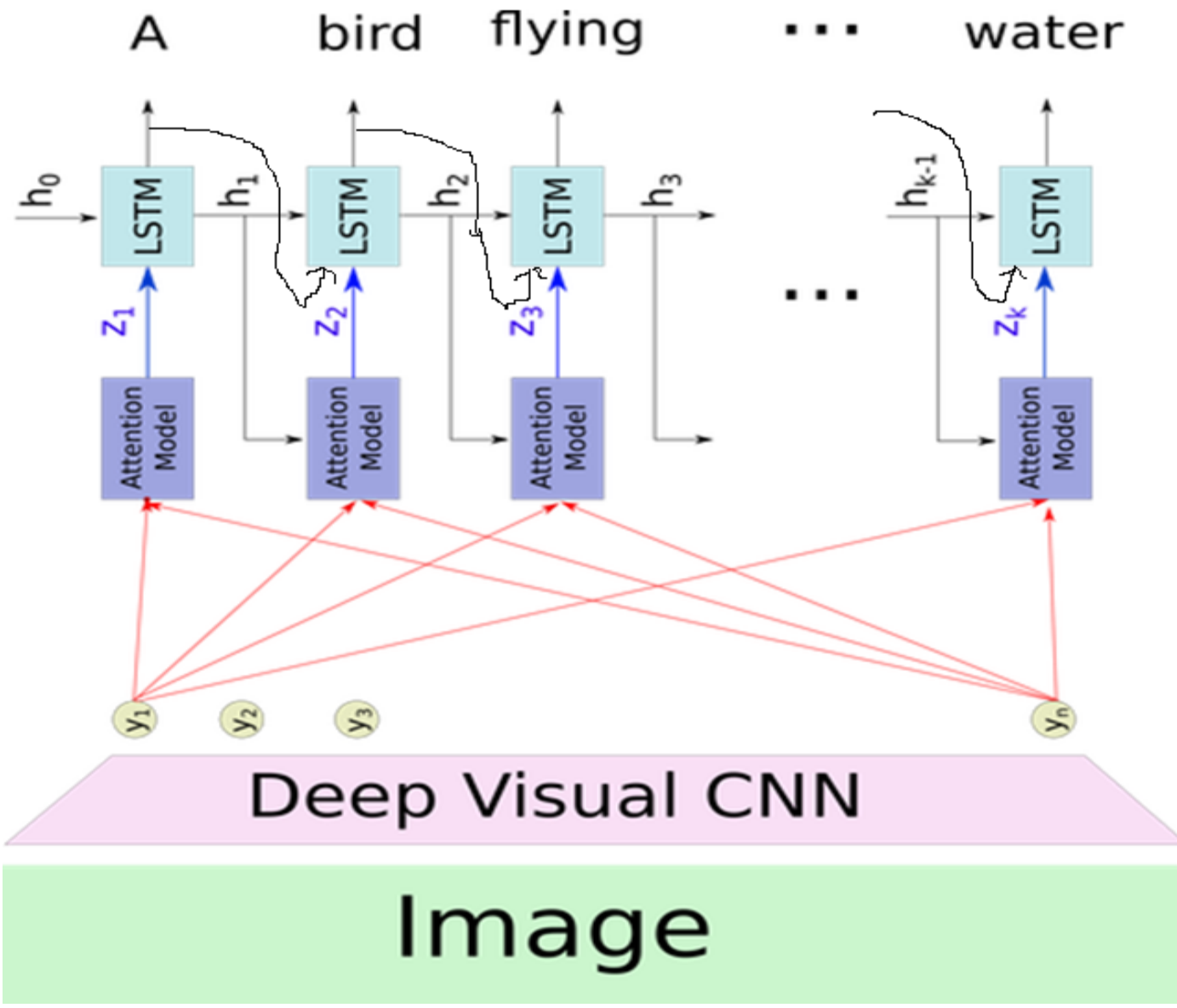
- The problem with this method is that, when the model is trying to generate the next word of the caption, this word is usually describing only a part of the image.

- Using the whole representation of the image to condition the generation of each word cannot efficiently produce different words for different parts of the image. This is exactly where an attention mechanism is helpful.

Attention based solution

- With an attention mechanism, the image is first divided into parts, and we compute with a Convolutional Neural Network (CNN) representations of each part .

- When the RNN is generating a new word, the attention mechanism is focusing on the relevant part of the image, so the decoder only uses specific parts of the image.


A
bird
flying
...
water
h_0
LSTM
h_1
LSTM
h_2
LSTM
h_3
h_{k-1}
LSTM
z_1
z_2
z_3
z_k
Attention Model
Attention Model
Attention Model
Attention Model
y_1
y_2
y_3
y_n
Deep Visual CNN
Image

- If we would have predicted i words, the hidden state of the LSTM is hi. We select the « relevant » part of the image by using hi as the context.

- The output of the attention model , which is the representation of the image filtered such that only the relevant parts of the image remains, is used as an input for the LSTM. Then, the LSTM predict a new word, and returns a new hidden state h i+1.

Image Caption Generation

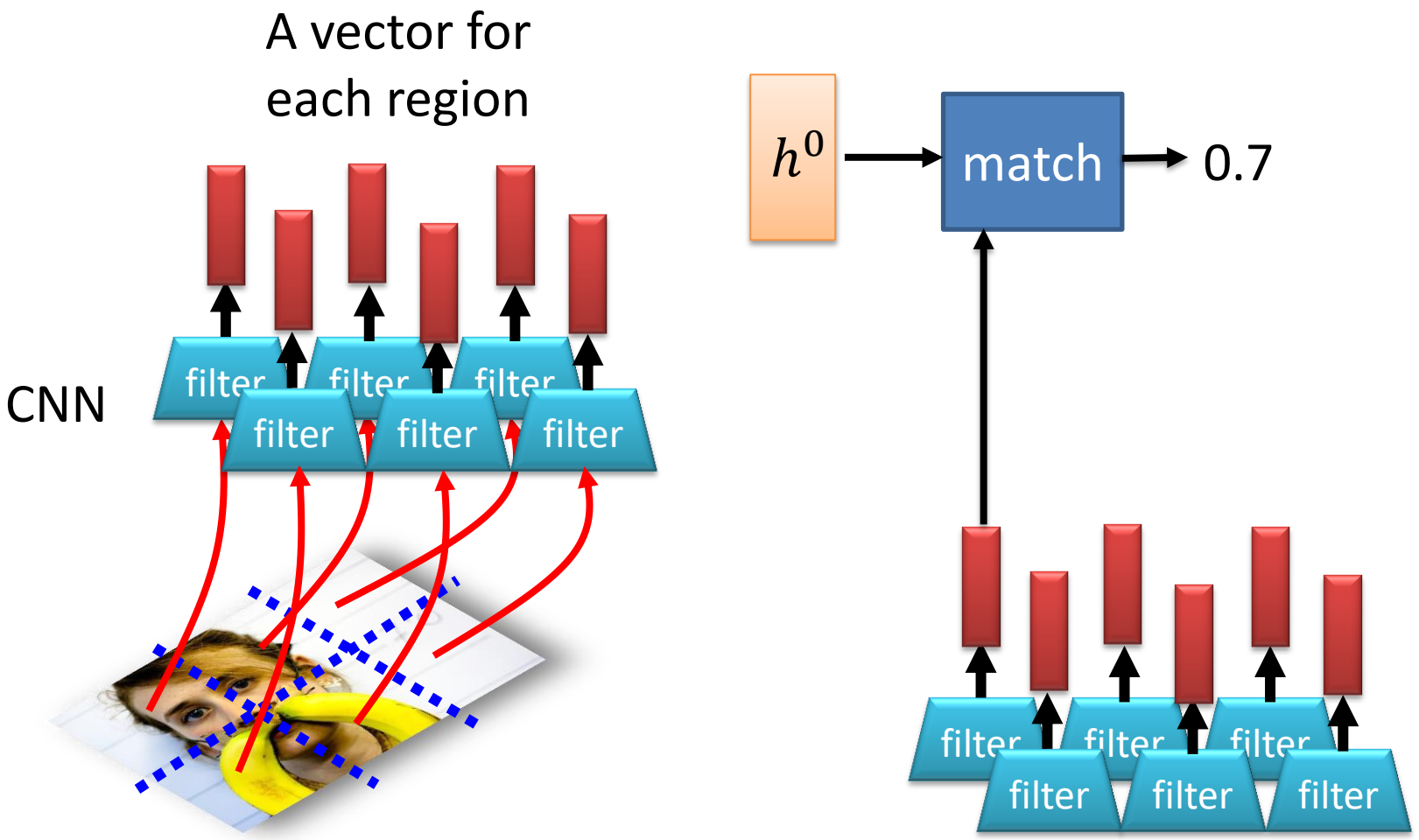


Image Caption Generation

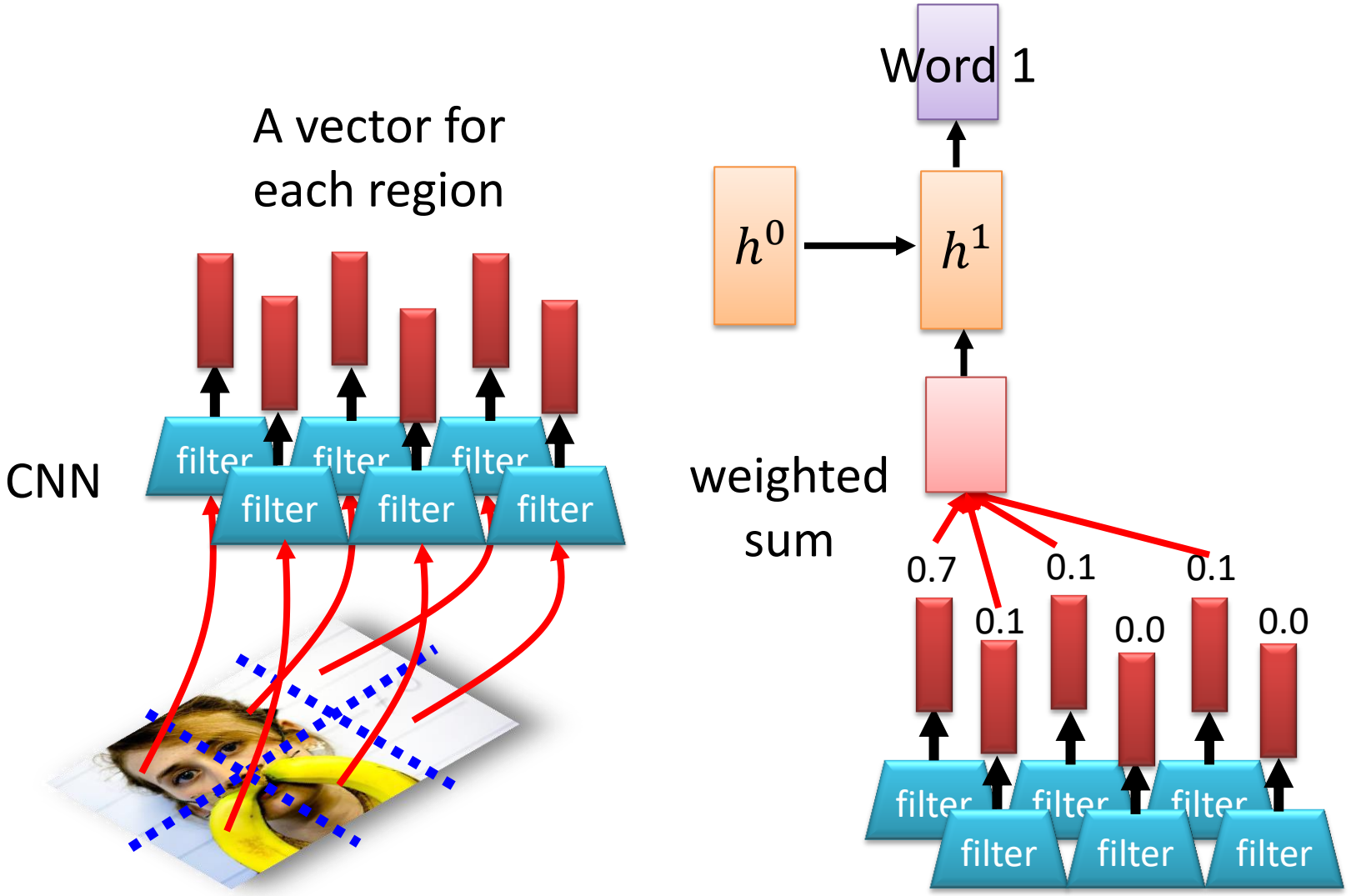


Image Caption Generation

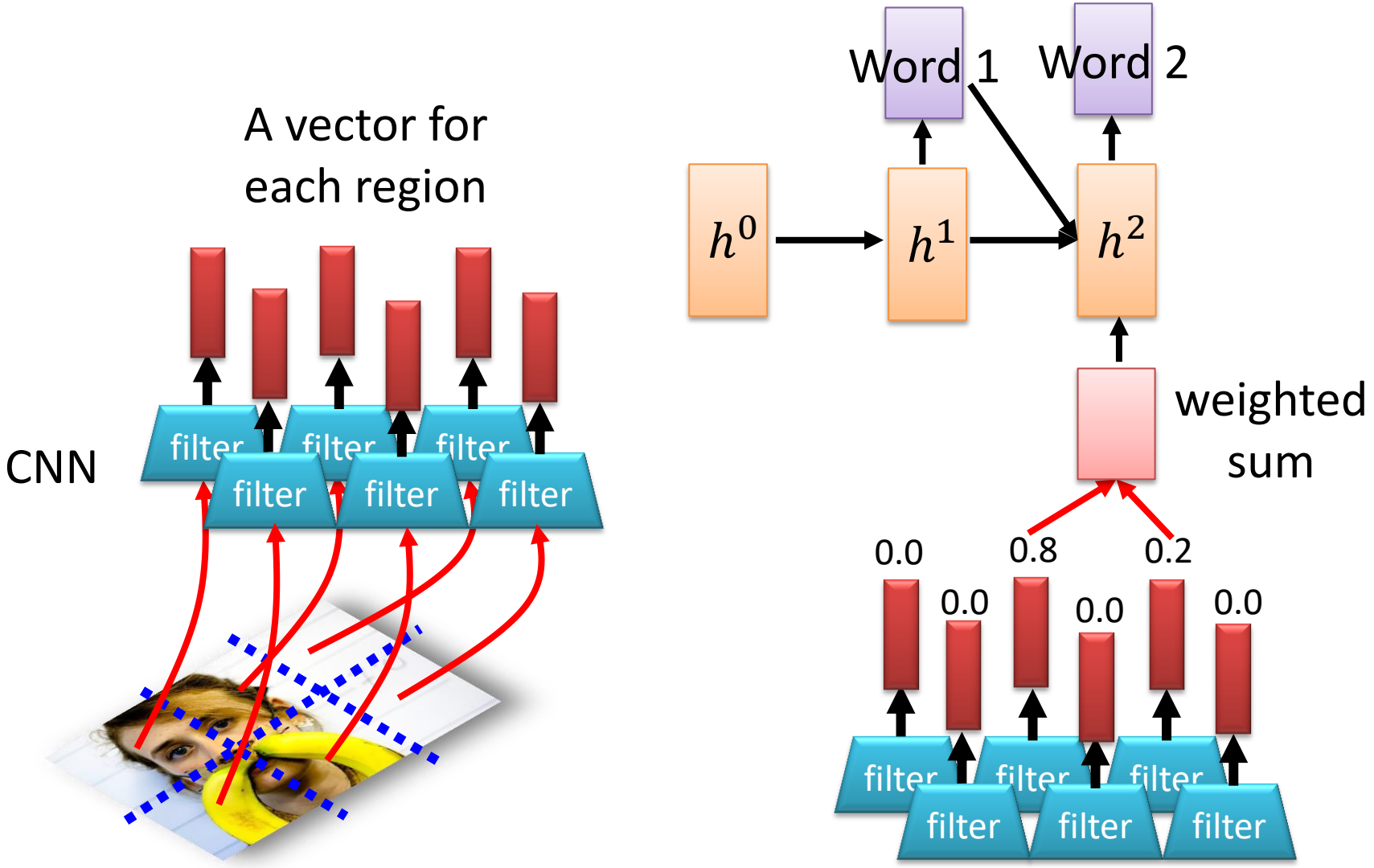


Image Caption Generation

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML, 2015

Image Caption Generation

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML, 2015

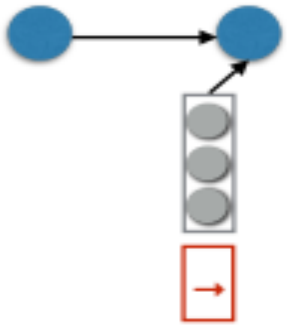
Video Caption Generation

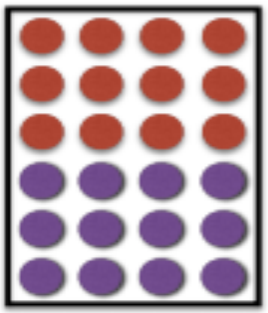
Video

A group of people is knocked by a tree.

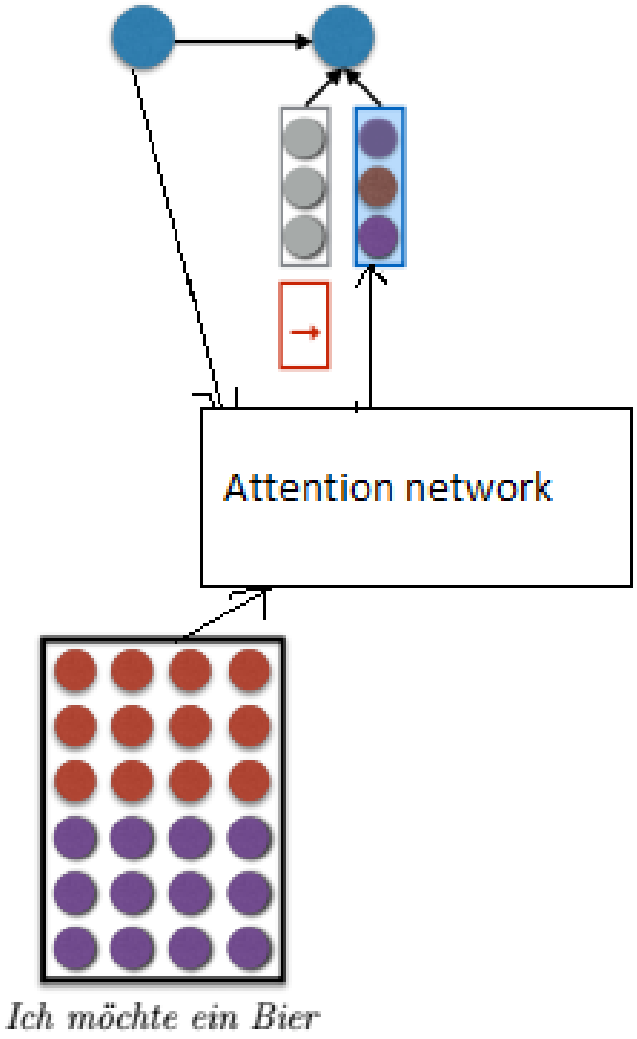
A group of people is walking in the forest.

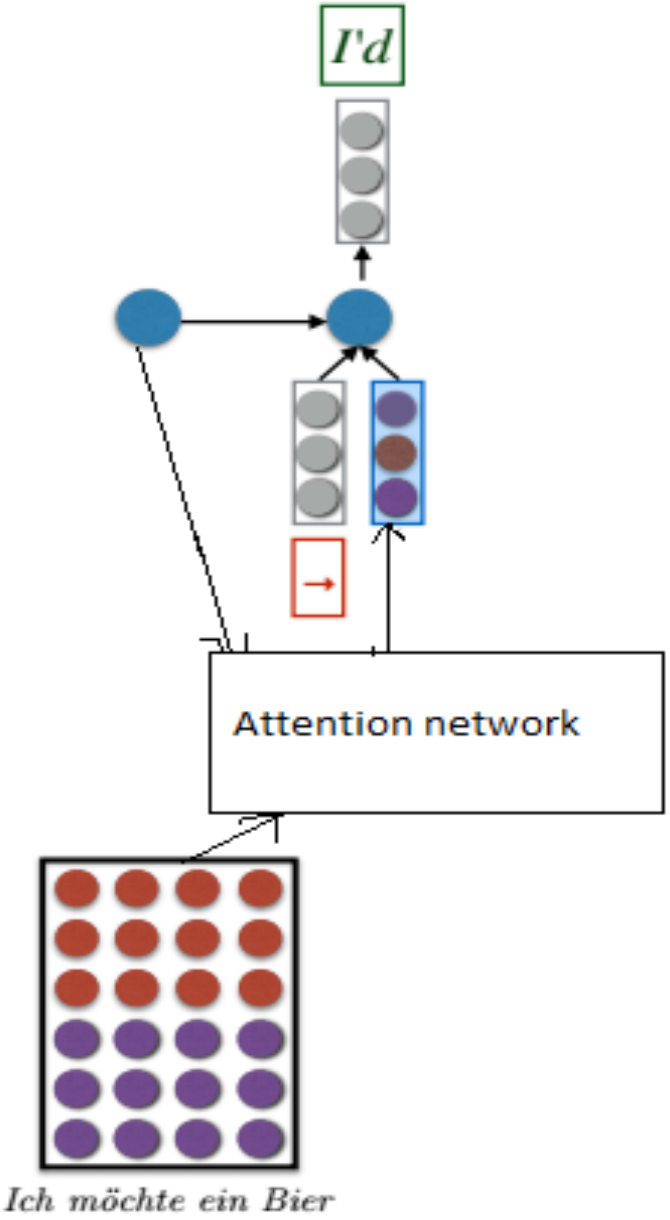
Decoder with Attention

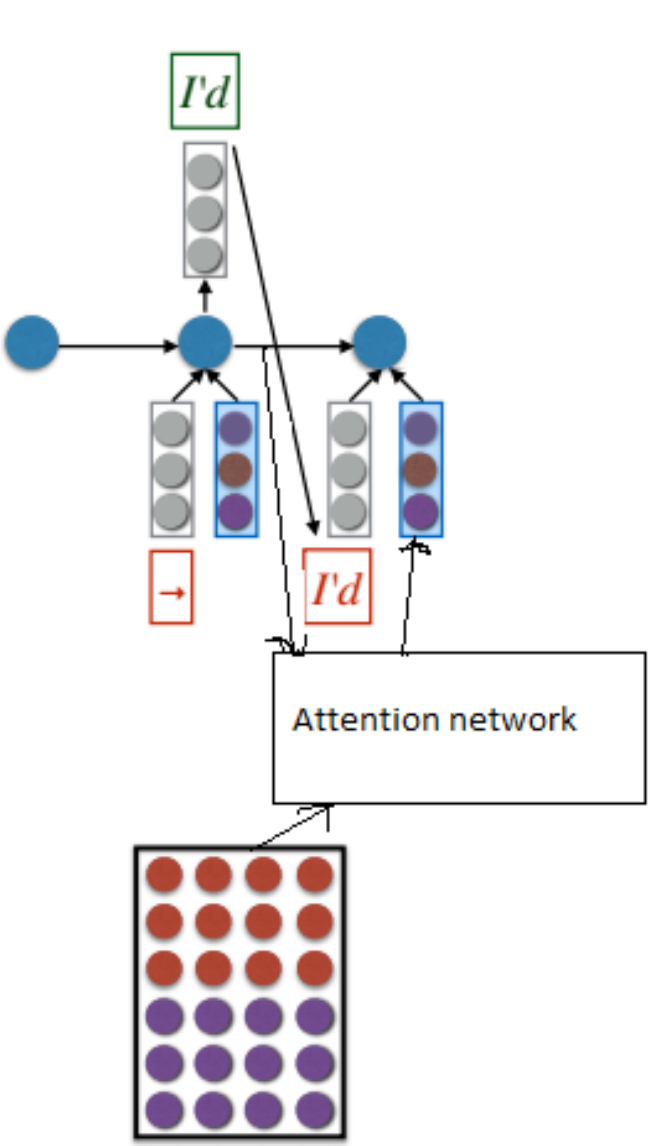


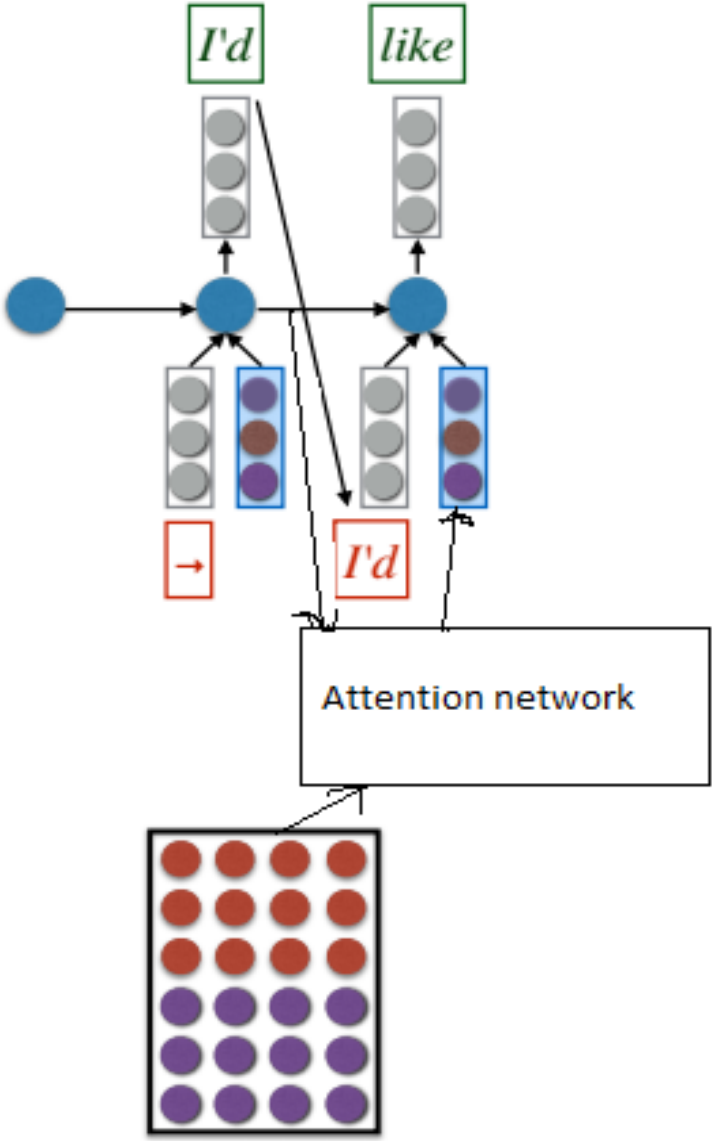


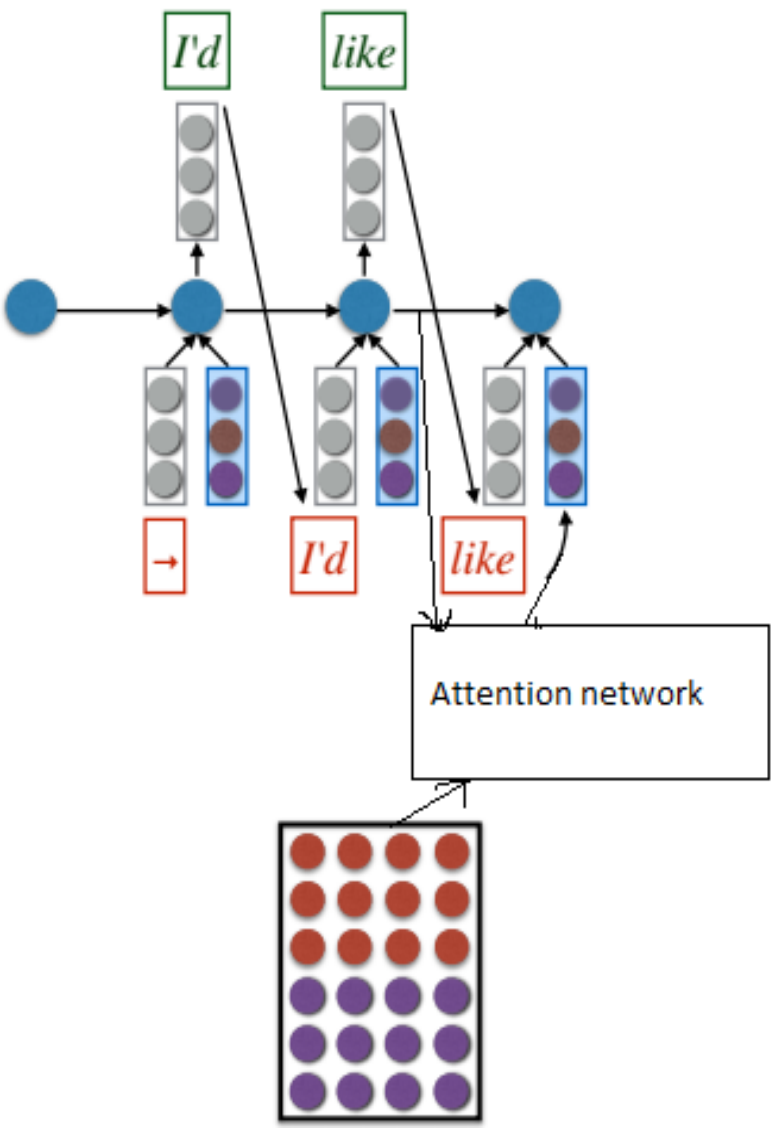
Ich möchte ein Bier

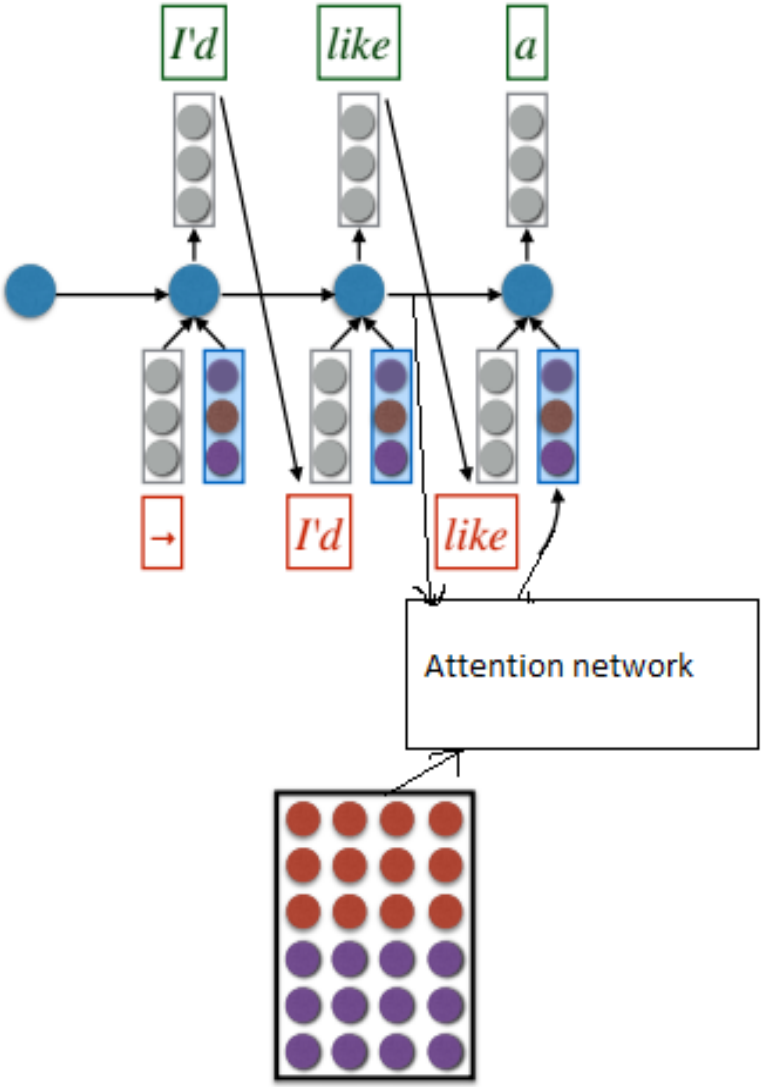

Attention network
Ich möchte ein Bier

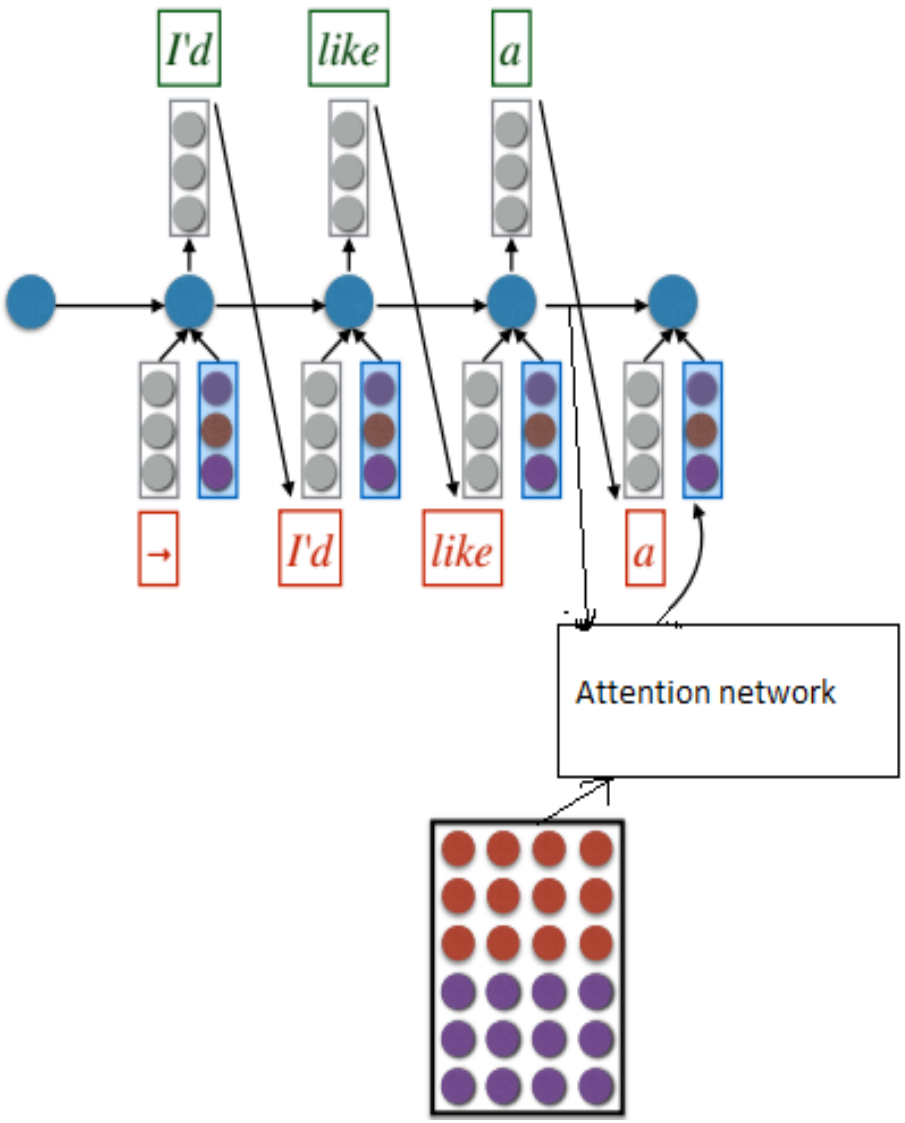

I'd
Attention network
Ich möchte ein Bier

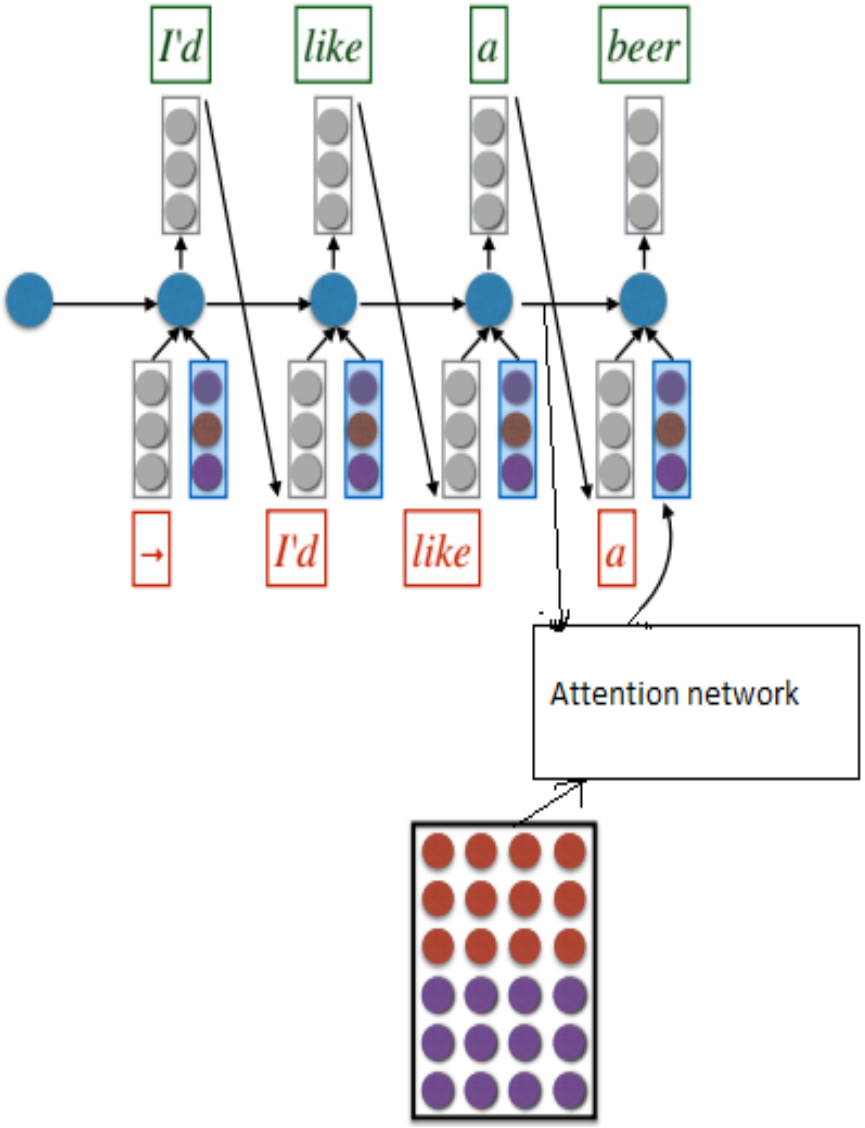

I'd
→
I'd
Attention network

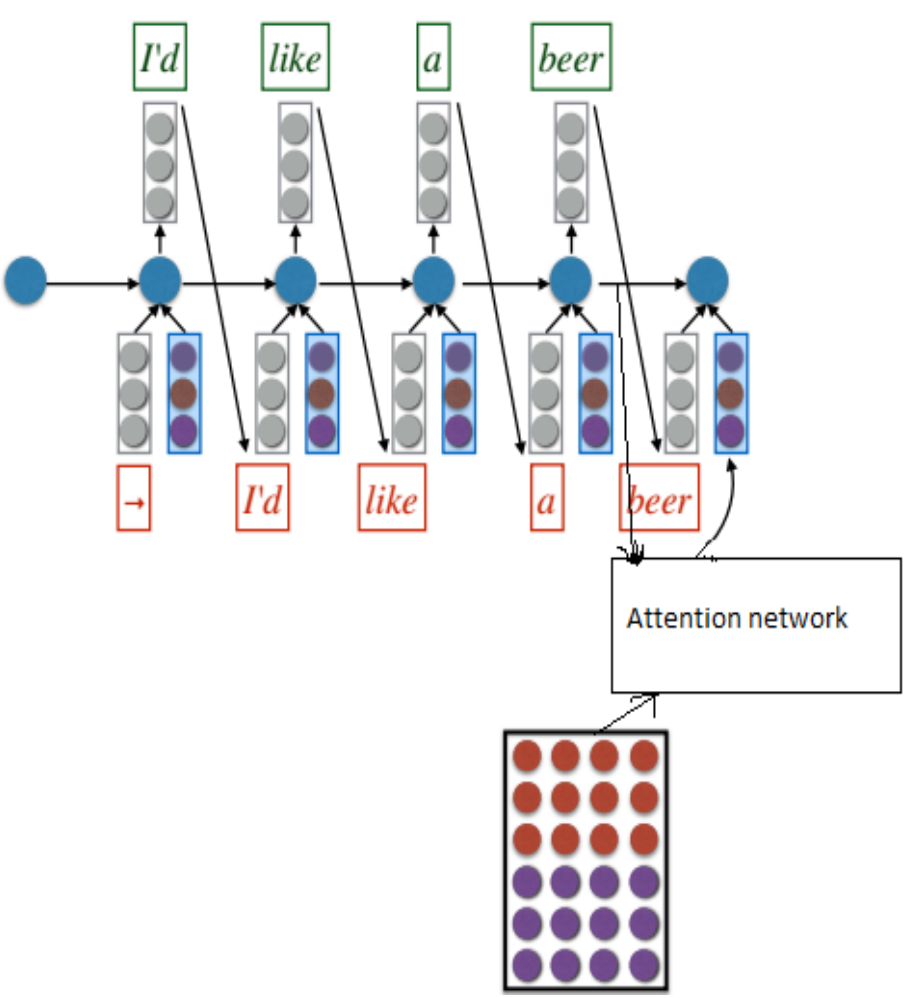

I'd
like
→
I'd
Attention network

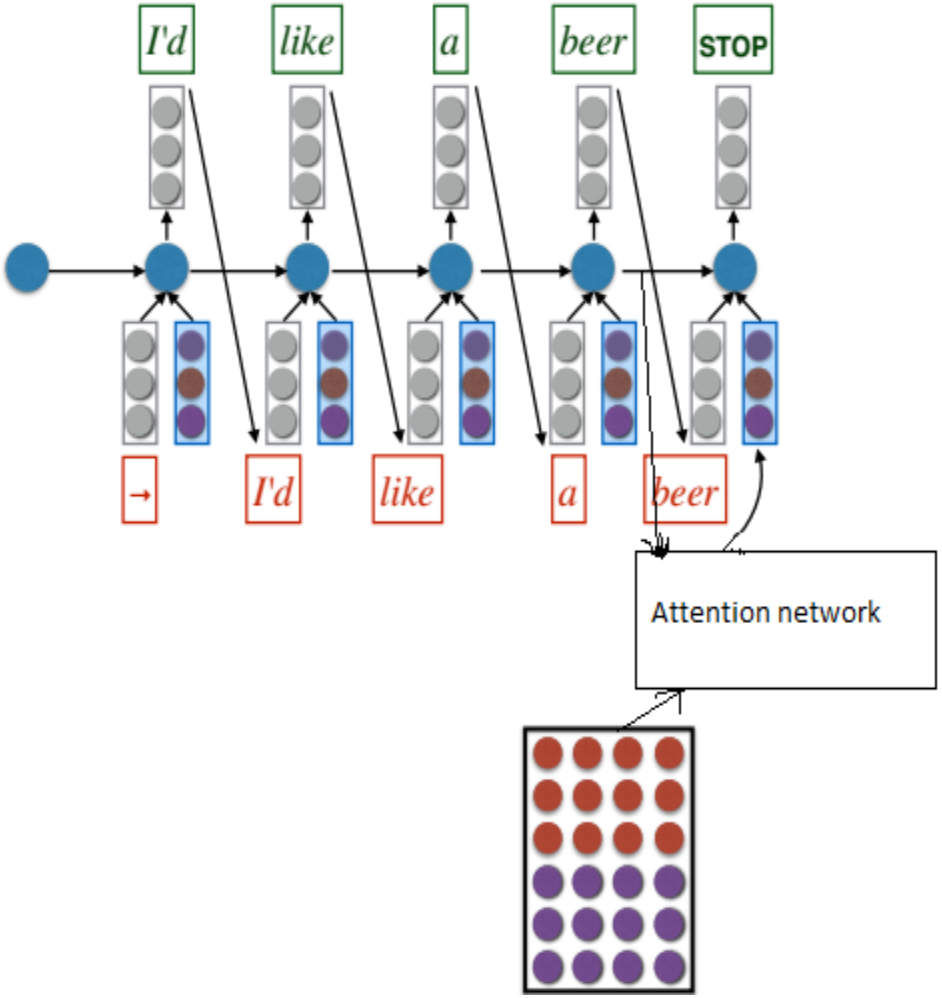

I'd
like
→
I'd
like
Attention network


I'd
like
a
→
I'd
like
Attention network


I'd
like
a
→
I'd
like
a
Attention network


I'd
like
a
beer
→
I'd
like
a
Attention network


I'd
like
a
beer
→
I'd
like
a
beer
Attention network


I'd
like
a
beer
STOP
→
I'd
like
a
beer
Attention network