CMPT 983

Grounded Natural Language Understanding

January 28, 2021

Attention

Today

- Attention
  - Review of attention mechanism
  - Types of attentions

Attention

Need for "attention"

- Uses encoding of entire input when generating each output token
- Maybe would be useful to focus on a part of the input as the output tokens are generated

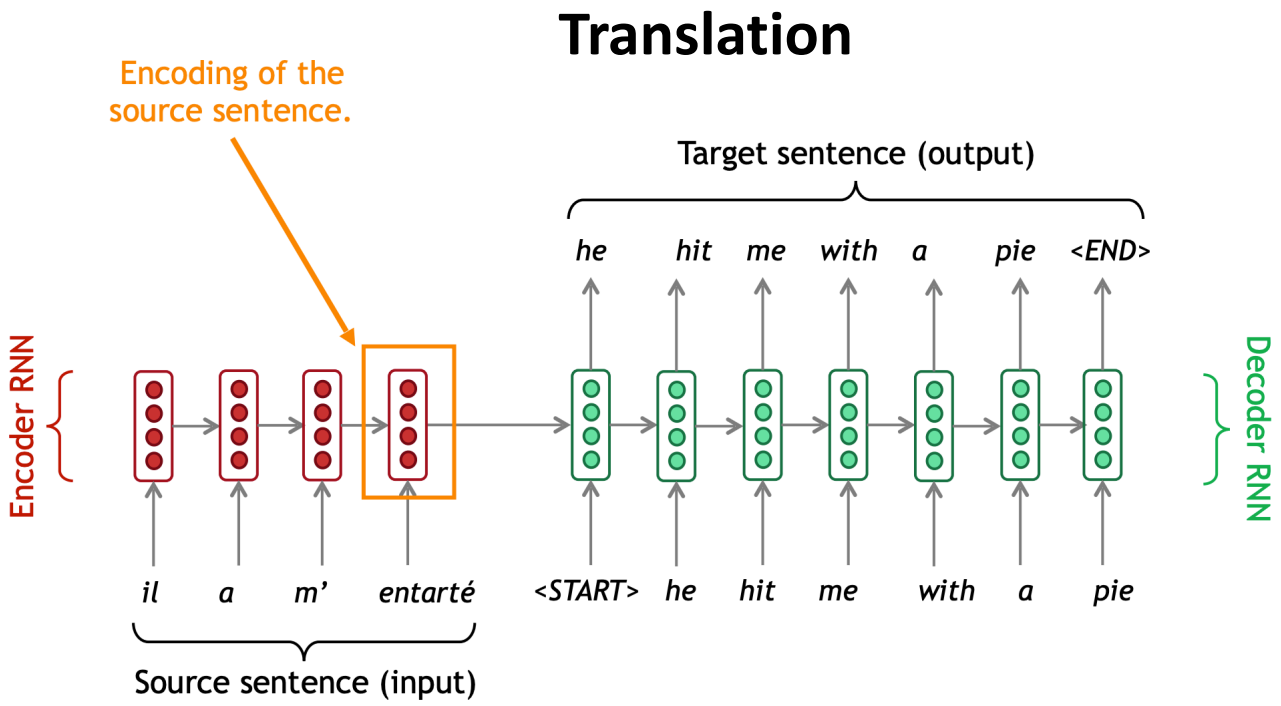


Image credit: Abigail See

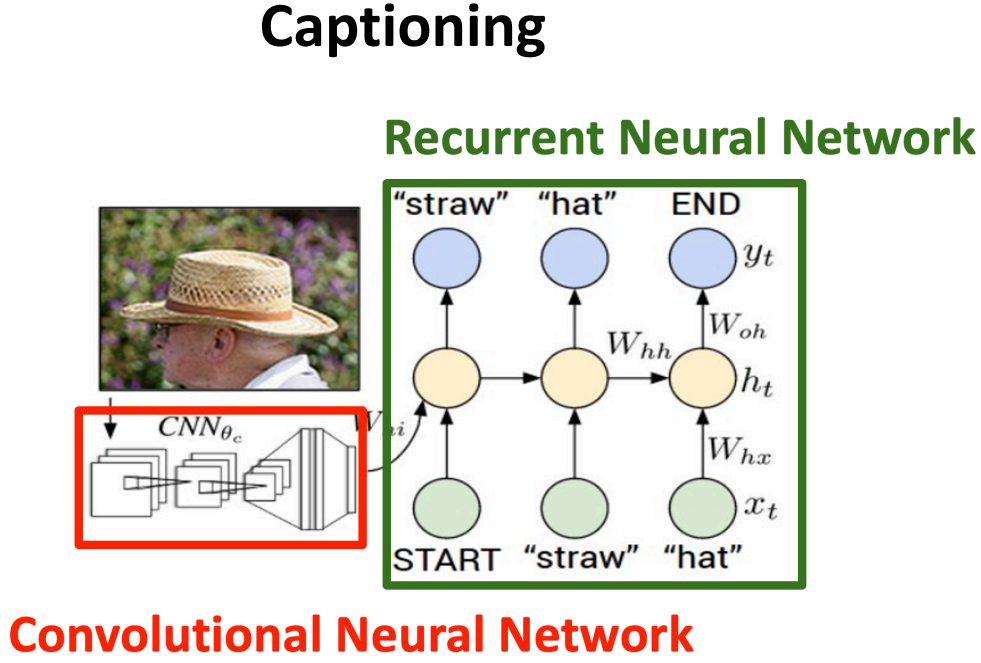


Image credit: Andrej Karpathy

Attention for VLN

Not every part of an input is important given the task context

Exit the bathroom. Turn left and exit the room using the door on the left. Wait there.

Slide credit: Stefan Lee

Attention for VLN

Not every part of an input is important given the task context

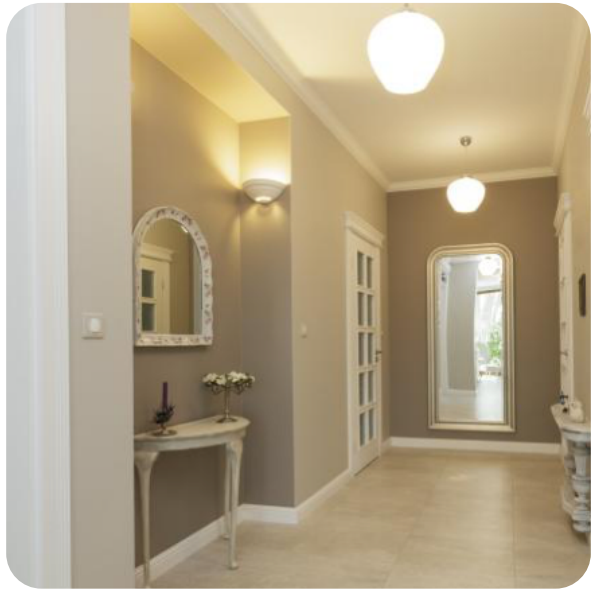
Exit the bathroom. Turn left and exit the room using the door on the left. Wait there.

Slide credit: Stefan Lee

Attention for VLN

Not every part of an input is important given the task context

Exit the bathroom. Turn left and exit the room using the door on the left. Wait there.



Slide credit: Stefan Lee

Attention for VLN

Not every part of an input is important given the task context

Take a right when you see the mirrored wardrobe.

Slide credit: Stefan Lee

Attention for VQA

Not every part of an input is important given the task context

What shape is the doormat?

Slide credit: Stefan Lee

Attention

- The concept of ‘attention’ has seen widespread use…
  - In many language and / or vision tasks, attention works extremely well!
  - Attention improves interpretability of neural networks

Q: What room are they in?

A: kitchen

Focus on part of input

Slides Credit: Peter Anderson

Attention mechanisms - summary

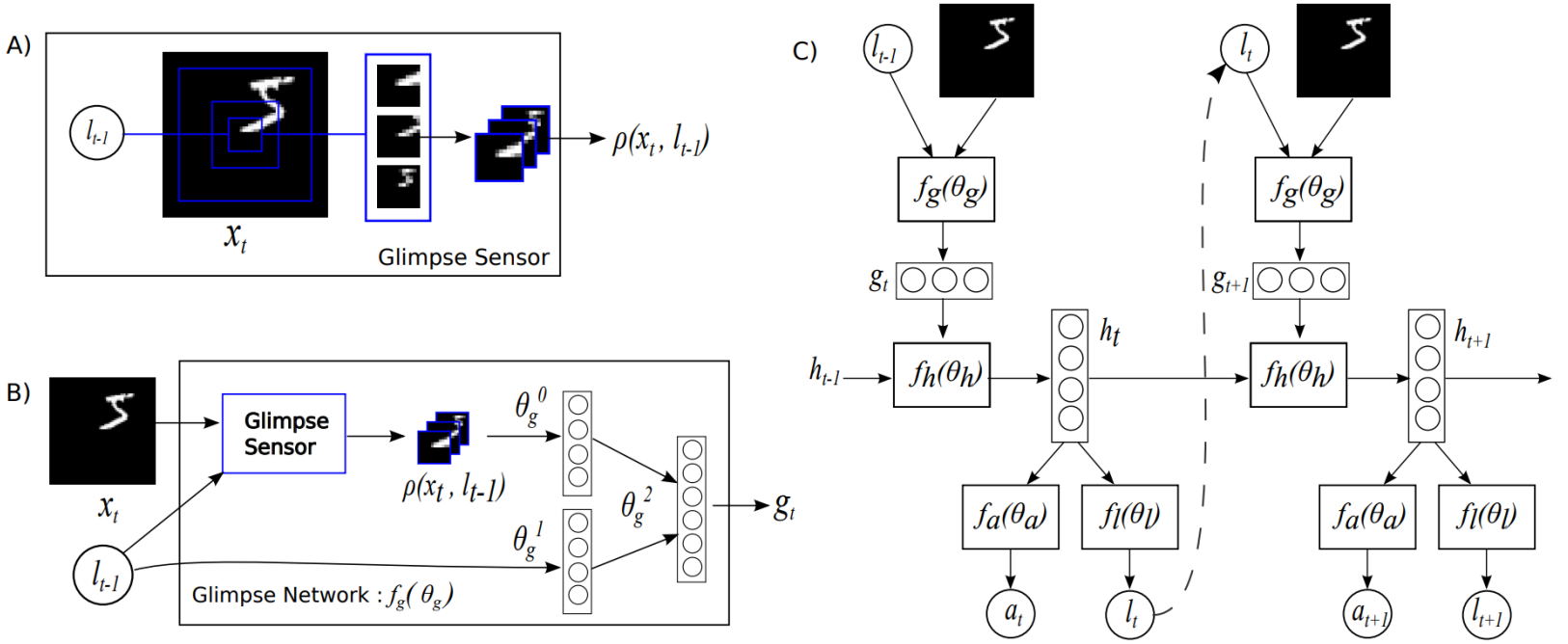
- Attention is probably one of the simplest and most effective ideas in deep learning – proven across many different domains
- Practically: focus on part of input by taking a weighted sum of different input parts
- With sufficient data, attention mechanisms can learn to ground language in visual content from ‘distant’ supervision
- Given the complexity of biological attention systems, assume there is still much to explore… particularly temporal aspects in context of LV&A

Slides Credit: Peter Anderson

Attention

- Major impact on computer vision and NLP from 2014/5

Recurrent Models of Visual Attention



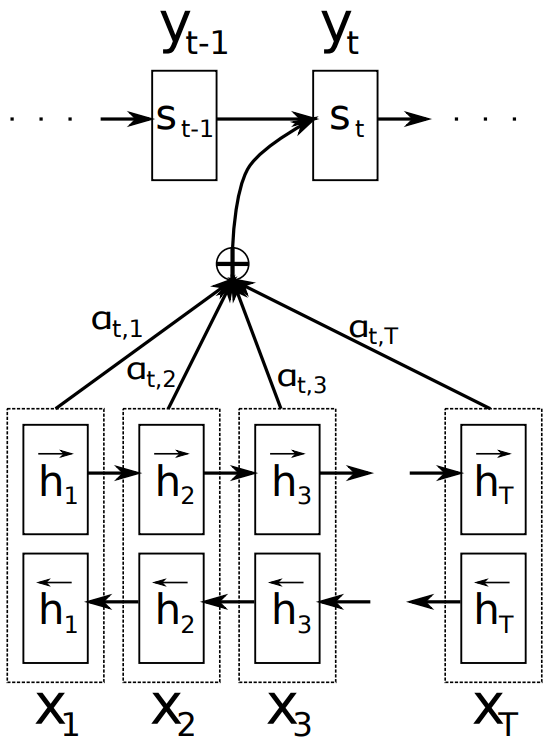
Mnih et al. NIPS 2014

Slides Credit: Peter Anderson

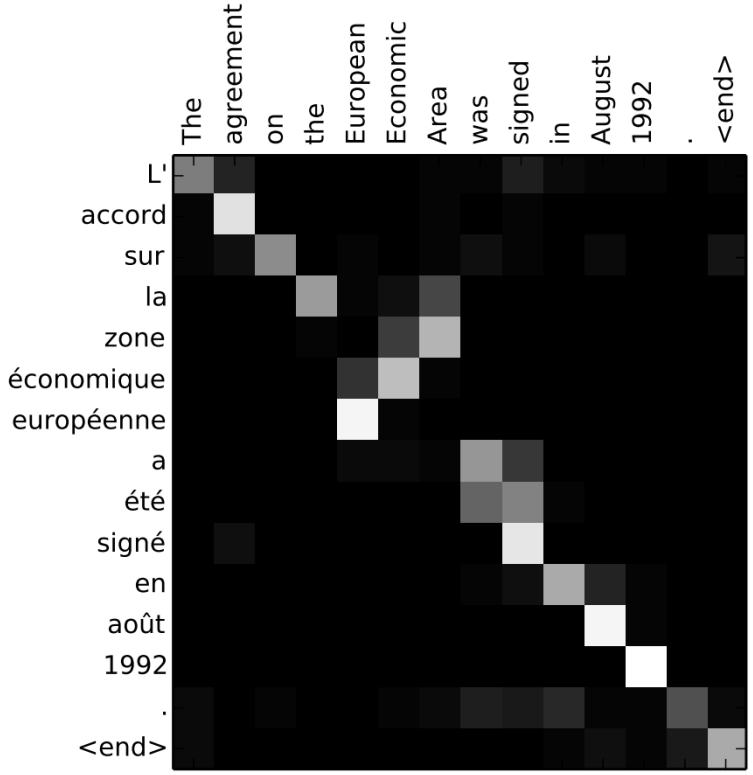
Attention

- Major impact on computer vision and NLP from 2014/5

Neural Machine Translation by Jointly Learning to Align and Translate



Bahdanau et al. ICLR 2015



generated (French)

source (English)

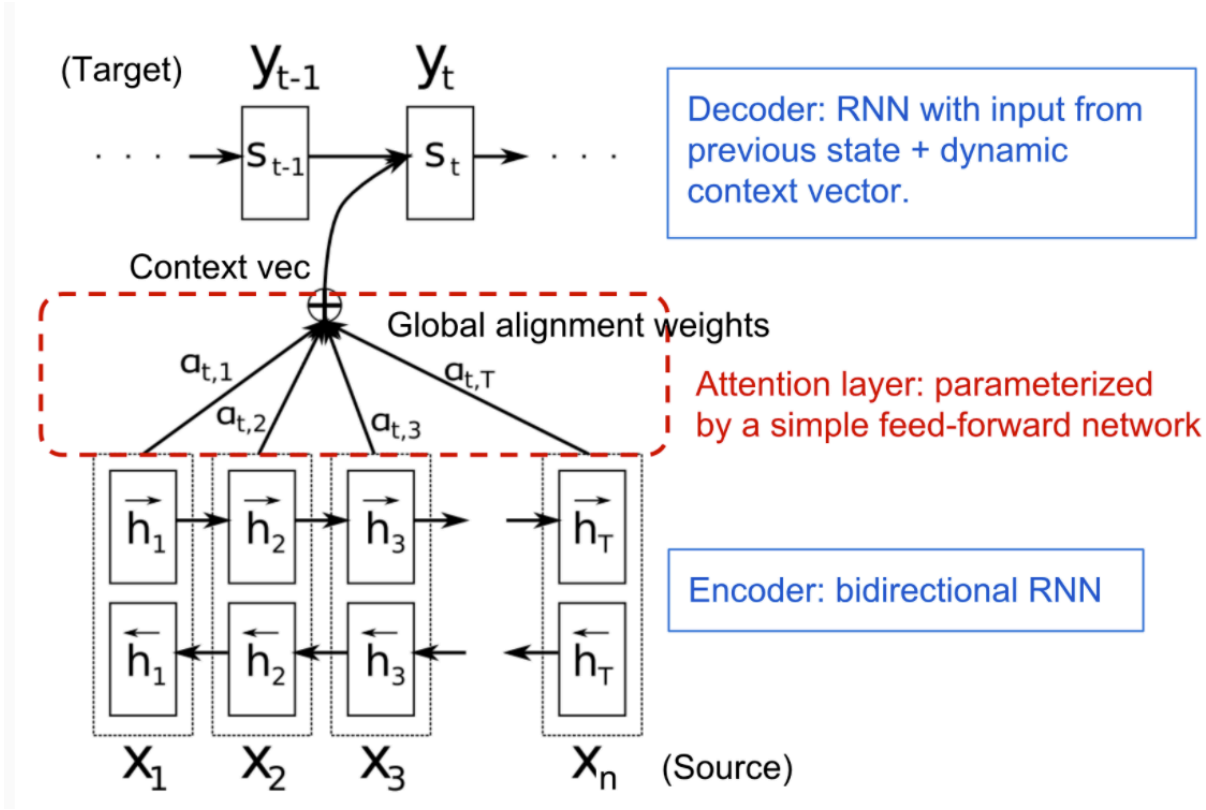
Attention weights α_i (0 = black, 1 = white)

Slides Credit: Peter Anderson

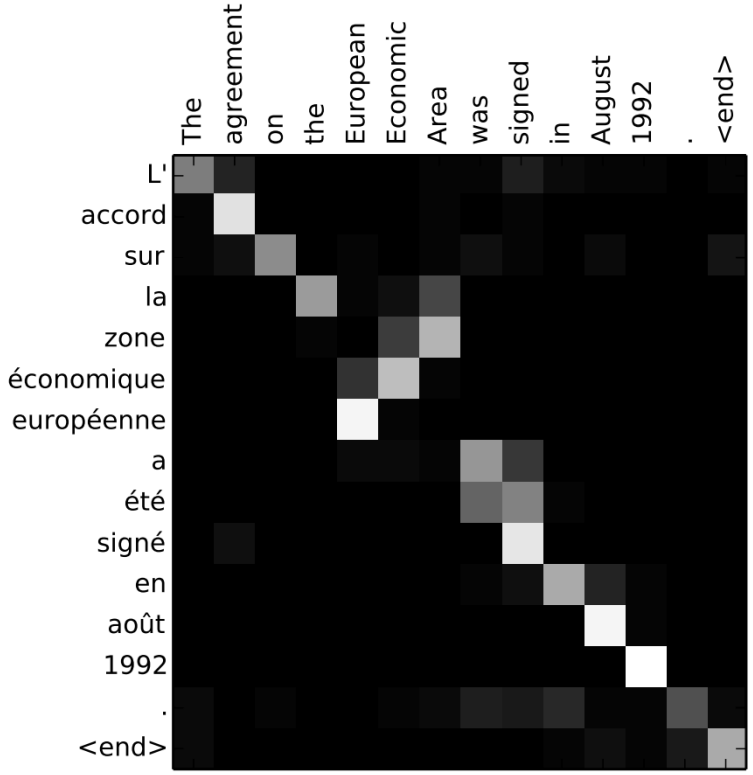
Attention

- Major impact on computer vision and NLP from 2014/5

Neural Machine Translation by Jointly Learning to Align and Translate



Bahdanau et al. ICLR 2015



generated (French)

source (English)

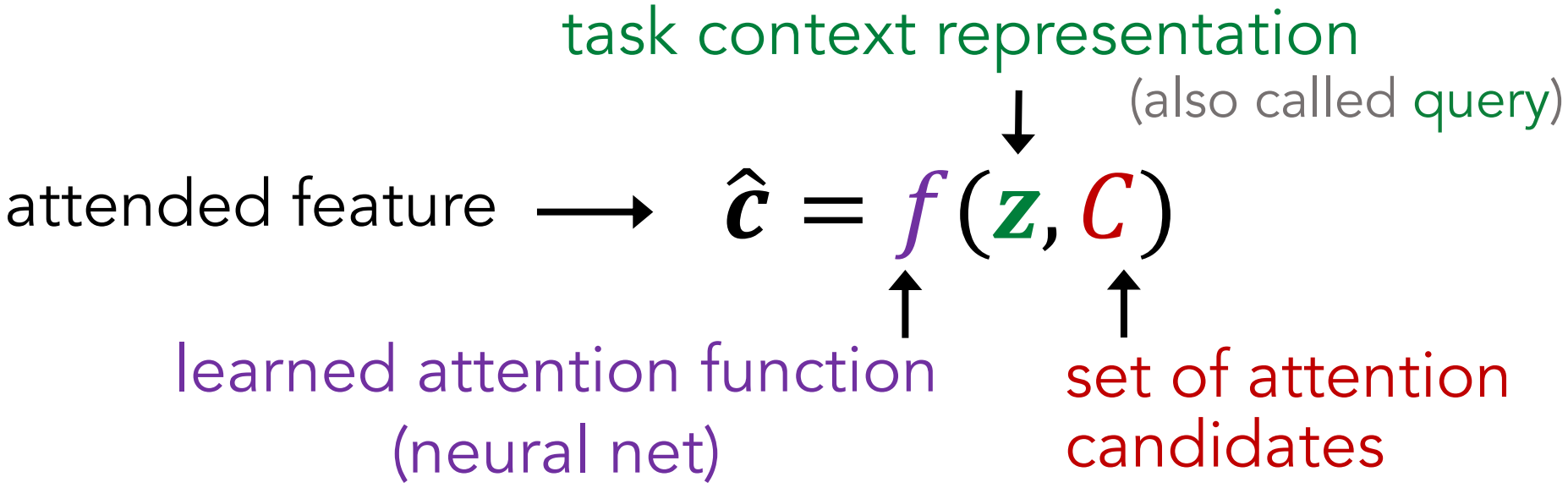
Attention weights α_i (0 = black, 1 = white)

Slides Credit: Peter Anderson

Attention

Attention in Neural Networks:

A learned mechanism that **learns to focus** on a subset of the **input** that is most **relevant to the current task**.



Slides Credit: Peter Anderson

Computing attention

Attention function, f

$$a_i = g(\boldsymbol{c}_i, \boldsymbol{z})$$
$$\boldsymbol{\alpha} = \text{softmax}(\boldsymbol{a})$$
$$\hat{\boldsymbol{c}} = \sum_{i=1}^{k} \alpha_i \, \boldsymbol{c}_i$$

Attention scores: $\boldsymbol{a}$ (unnormalized)

Attention weights: $\boldsymbol{\alpha}$ (normalized)

Final attention output

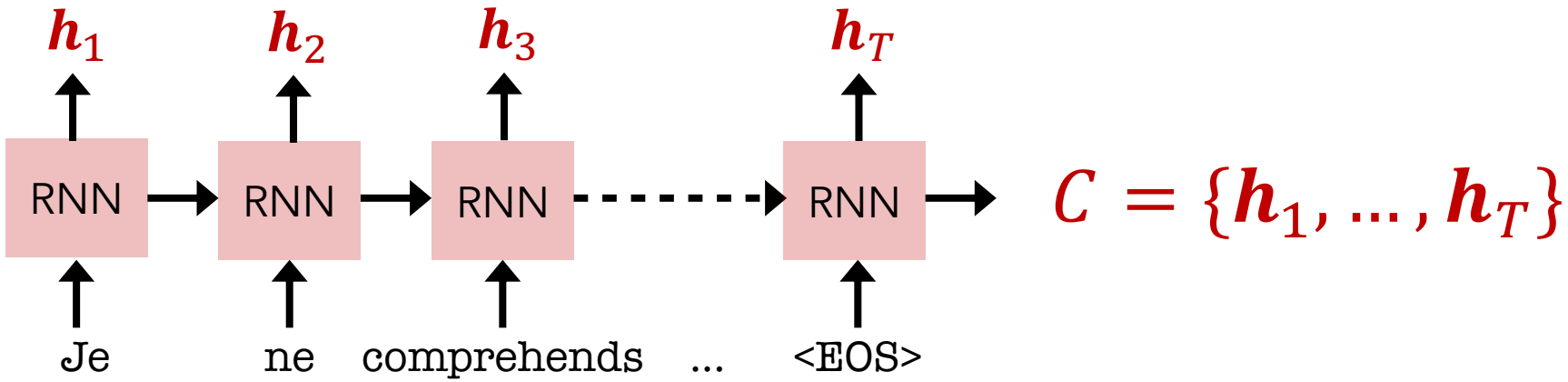
Weighted sum of context features

Attention score $a_i = g(\boldsymbol{c}_i, \boldsymbol{z})$
how well does the attention candidate $\boldsymbol{c}_i$ match the query $\boldsymbol{z}$

- Dot-product attention:
  $g(c_i, z) = z^\top c_i$
- Neural network
  $g(c_i, z) = v^\top \tanh\left(W_1 c_i + W_2 z\right)$

Attention over Text

Attention candidates, C typically defined by the hidden states of an encoder (e.g. one feature vector for each word in the input text)

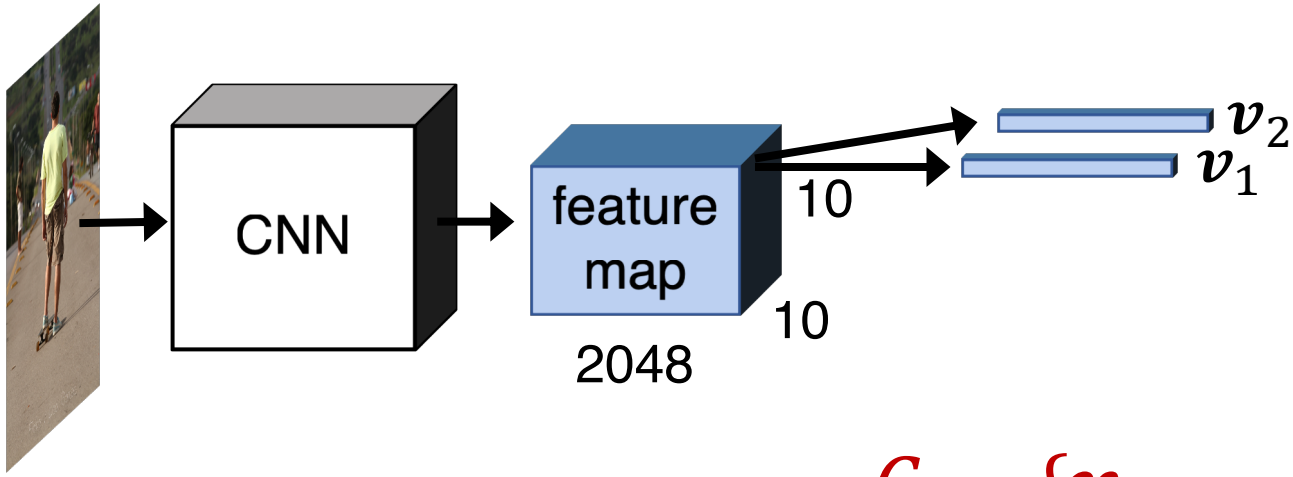


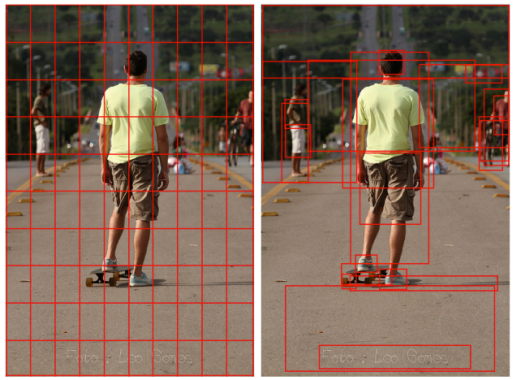
Slides Credit: Peter Anderson

Attention over Visual Features

- Attention candidates, $\boldsymbol{C}$ typically defined by the spatial output of a CNN (feature vectors for different parts of the image)

Grid based or over object proposals





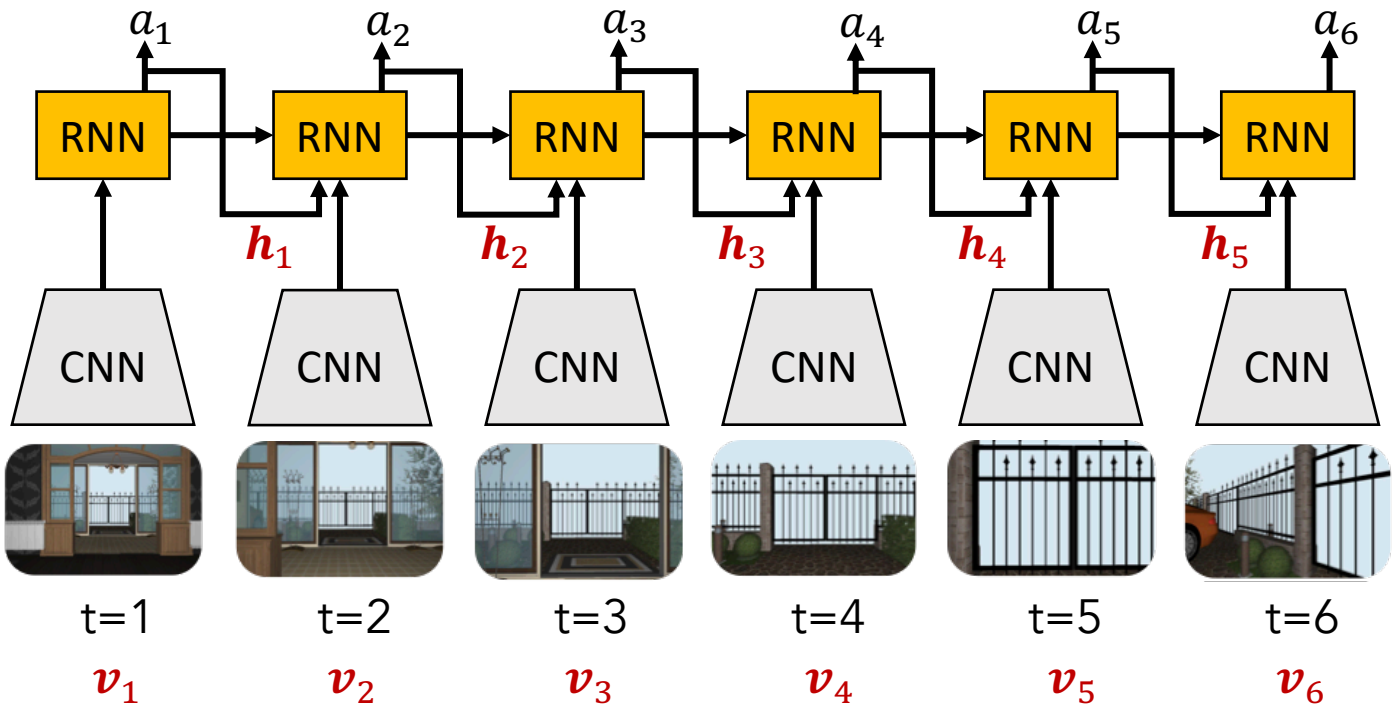
$$C = \{v_1, \ldots, v_{100}\}$$

Slides Credit: Peter Anderson

Attention over Agent Experience

Embodied AI (visual language navigation)

- Attention candidates, $\boldsymbol{C}$ as agent hidden state or visual vectors



$$C = \{h_1, \ldots, h_5\}$$

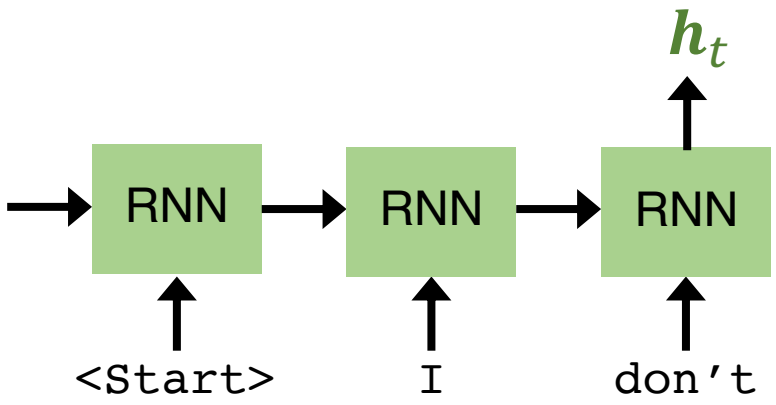
or

$$C = \{v_1, \ldots, v_6\}$$

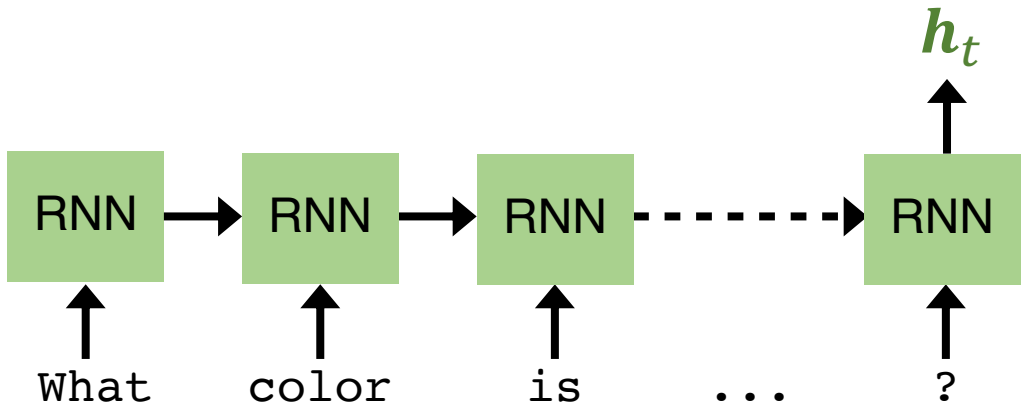
Task context for Attention

- Task context representation, $\boldsymbol{z}$, is often an RNN encoding

Machine translation / image captioning:
Decoder hidden state



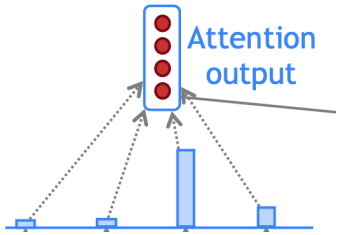
VQA: Question encoding
(final encoder hidden state)



Slides Credit: Peter Anderson

Putting it all together

- Attentive machine translation model



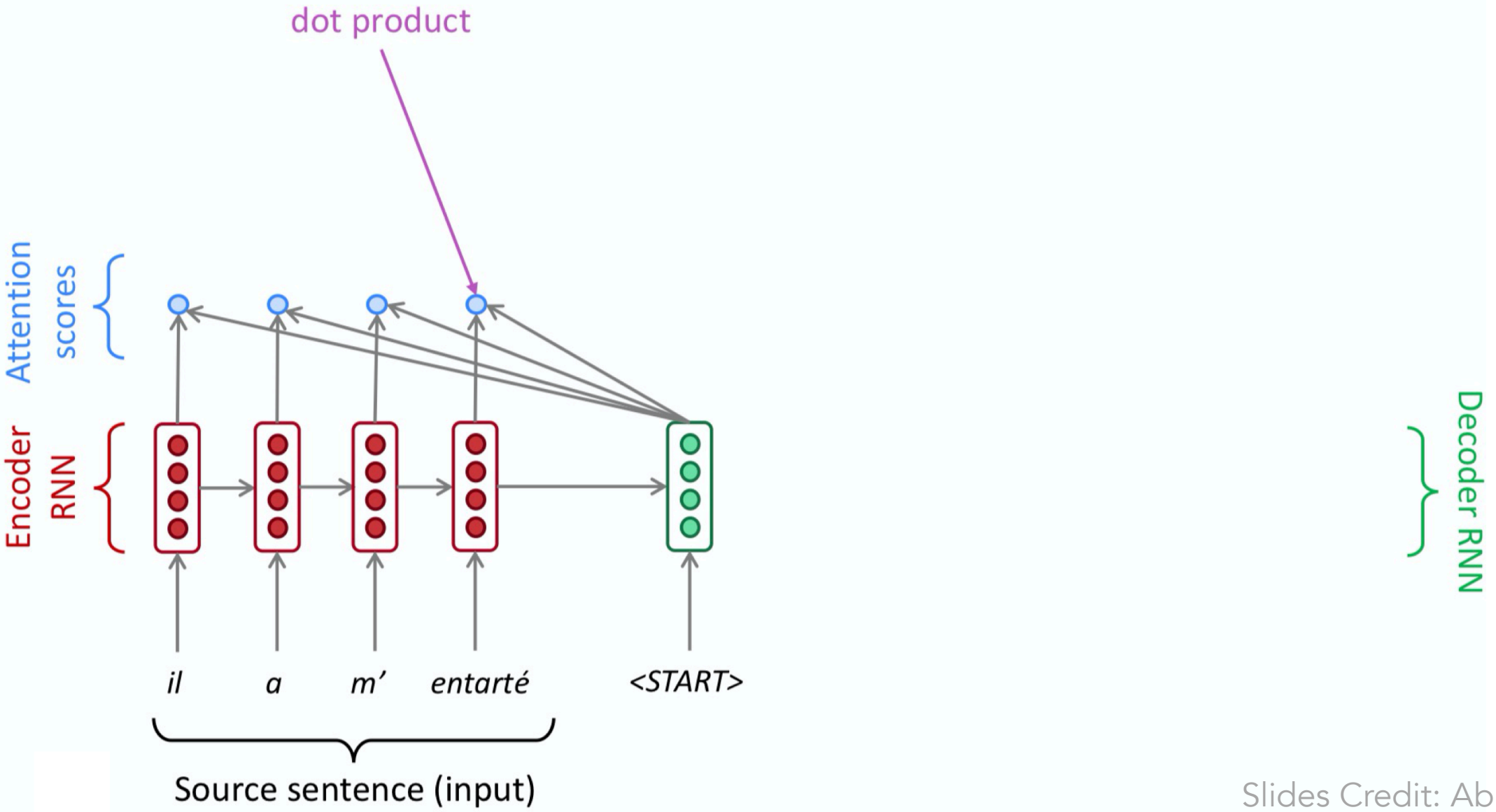
$\hat{\boldsymbol{c}}_t$ is the weighted sum of hidden states

$$C = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T\}$$

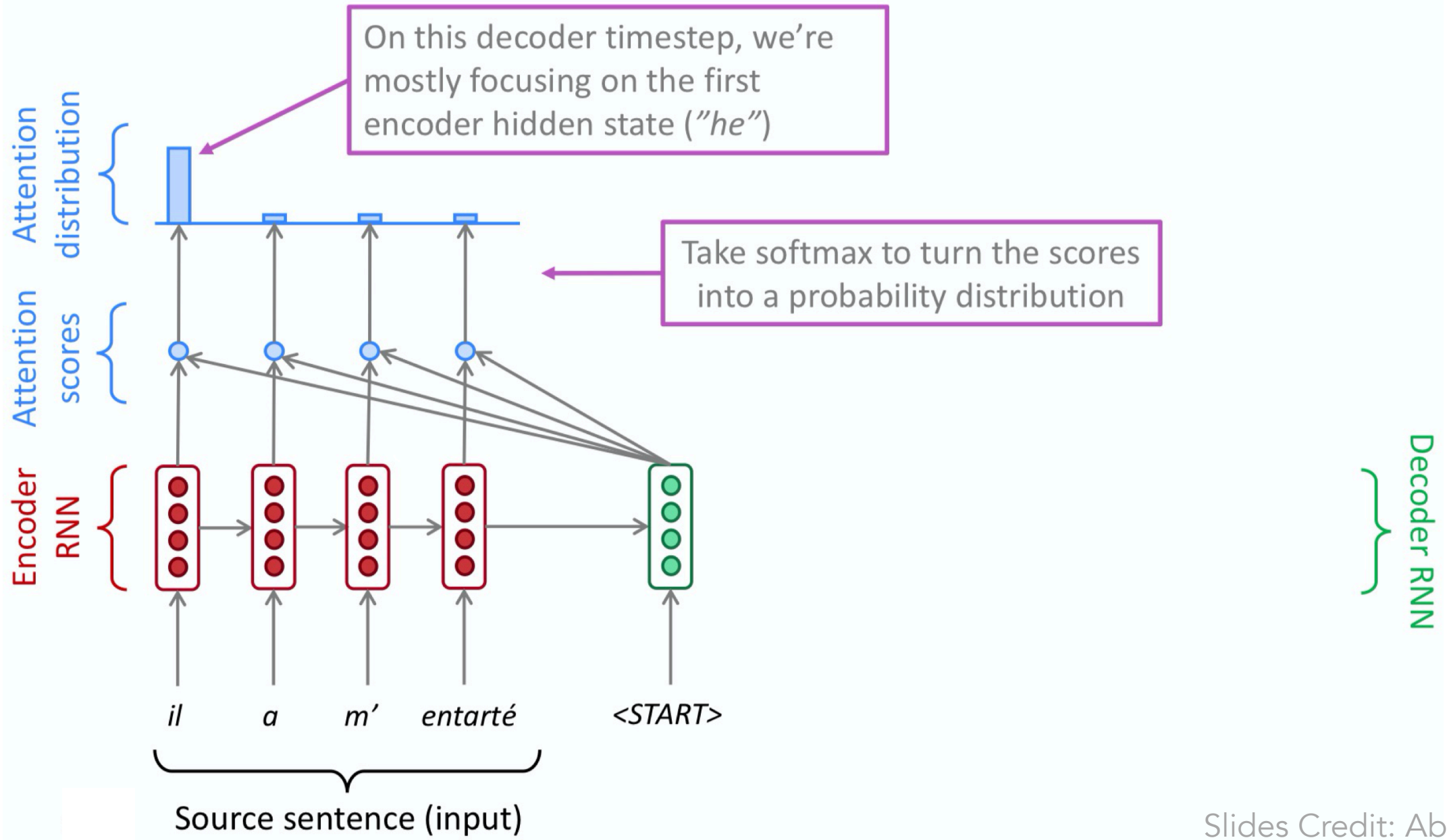
Slides Credit: Peter Anderson

Interlude: Attention for machine translation

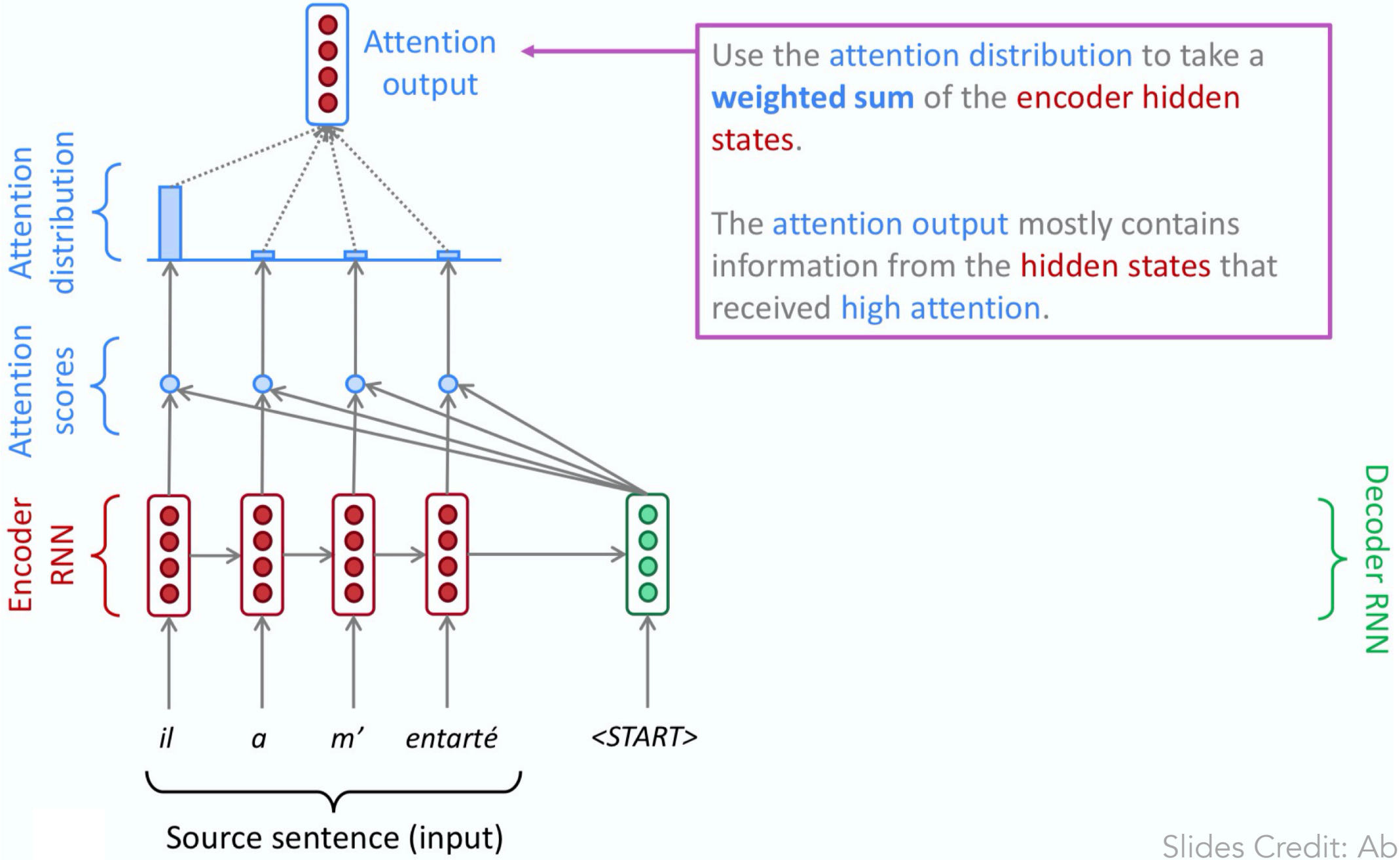
Attention for machine translation



Slides Credit: Abigail See



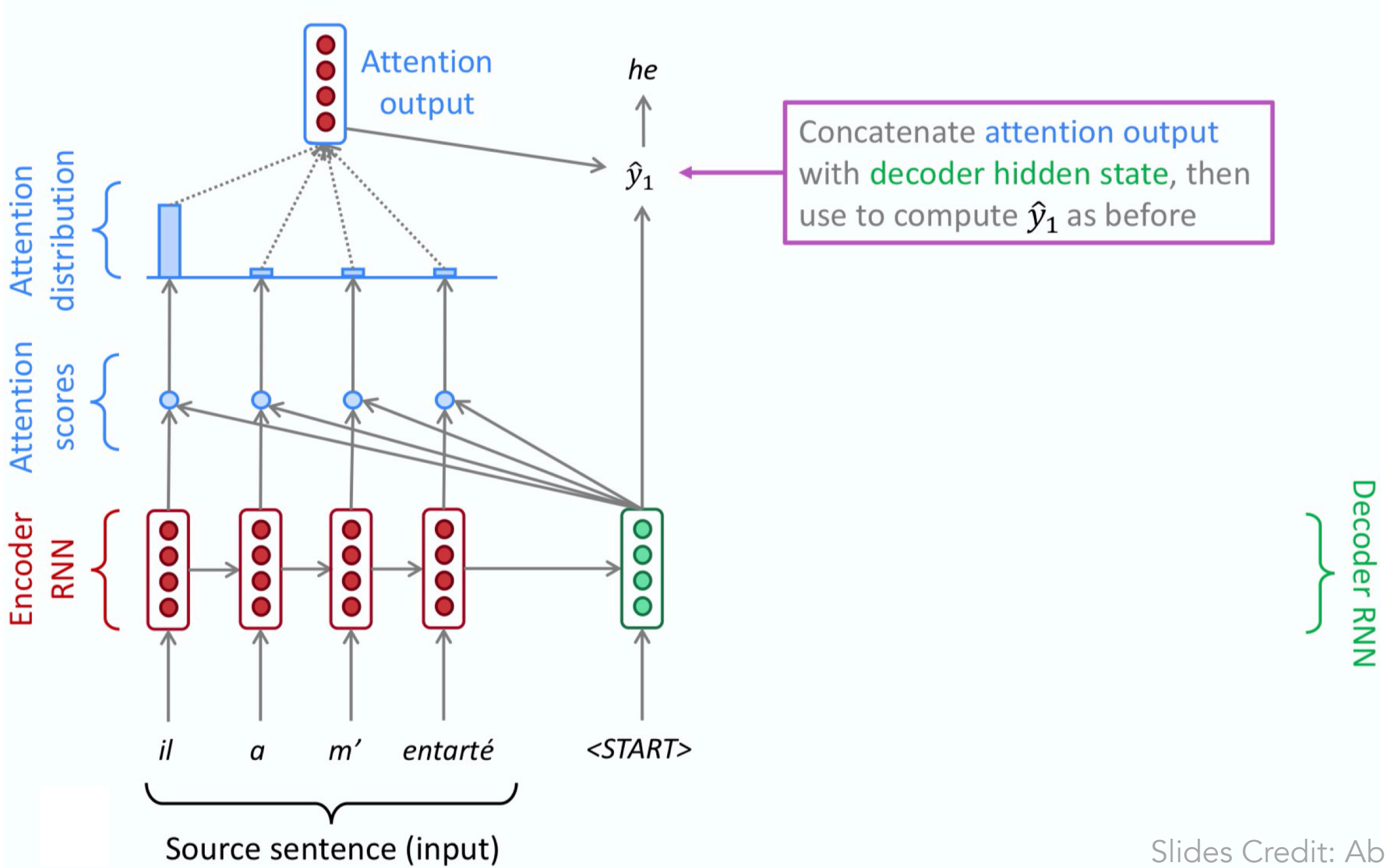
Slides Credit: Abigail See



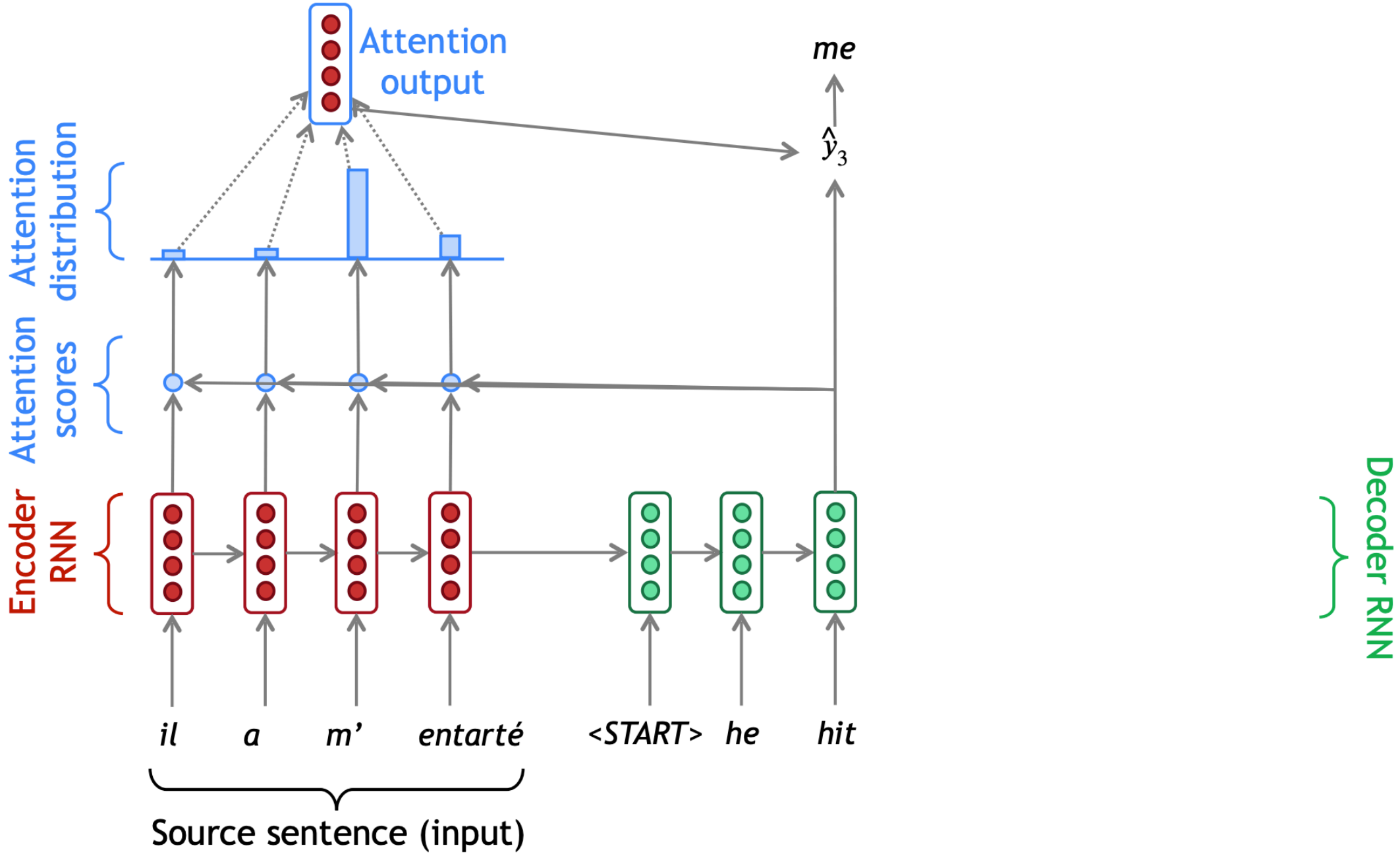
Use the attention distribution to take a **weighted sum** of the encoder hidden states.

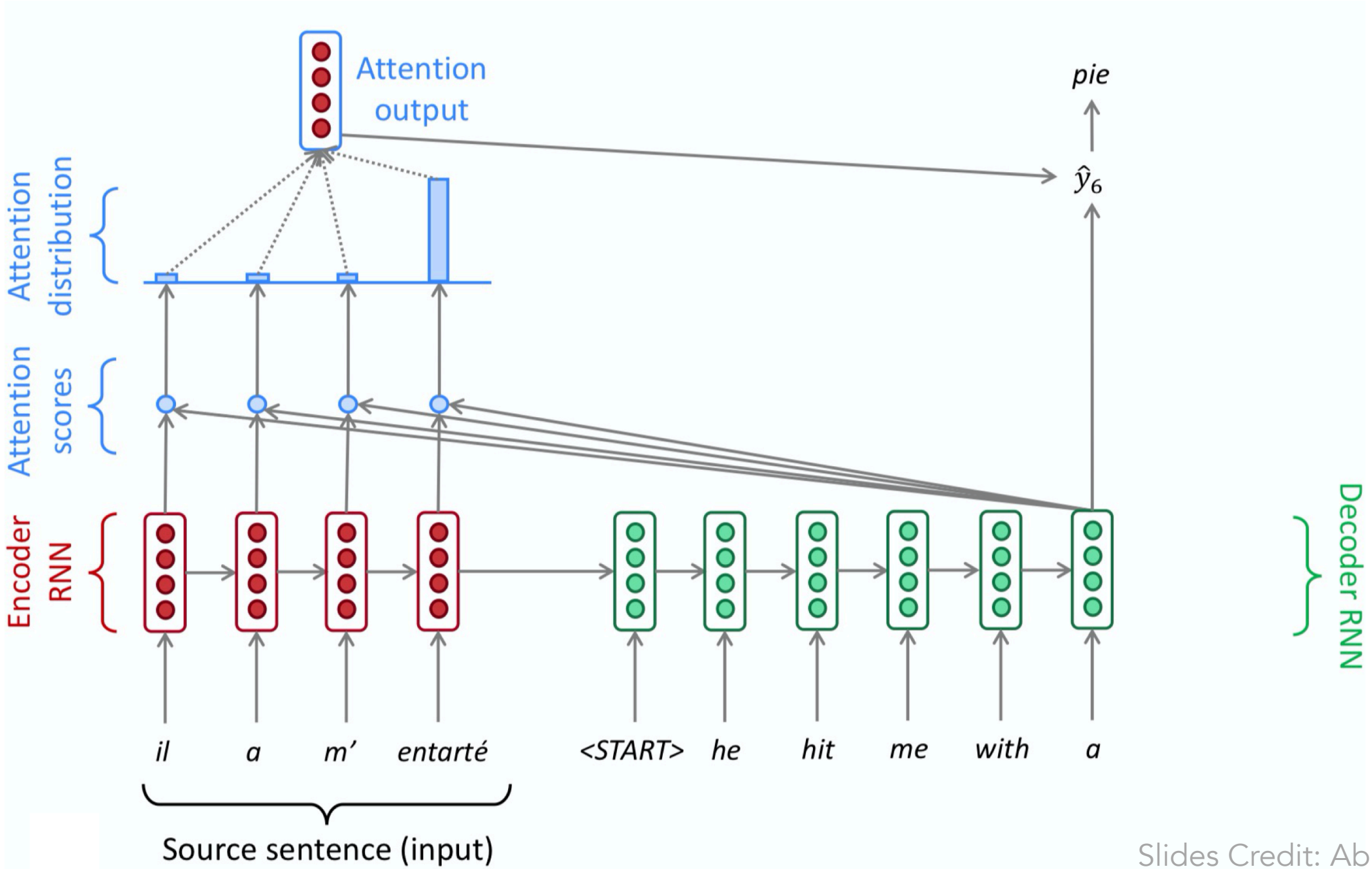
The attention output mostly contains information from the hidden states that received high attention.

Slides Credit: Abigail See



Slides Credit: Abigail See

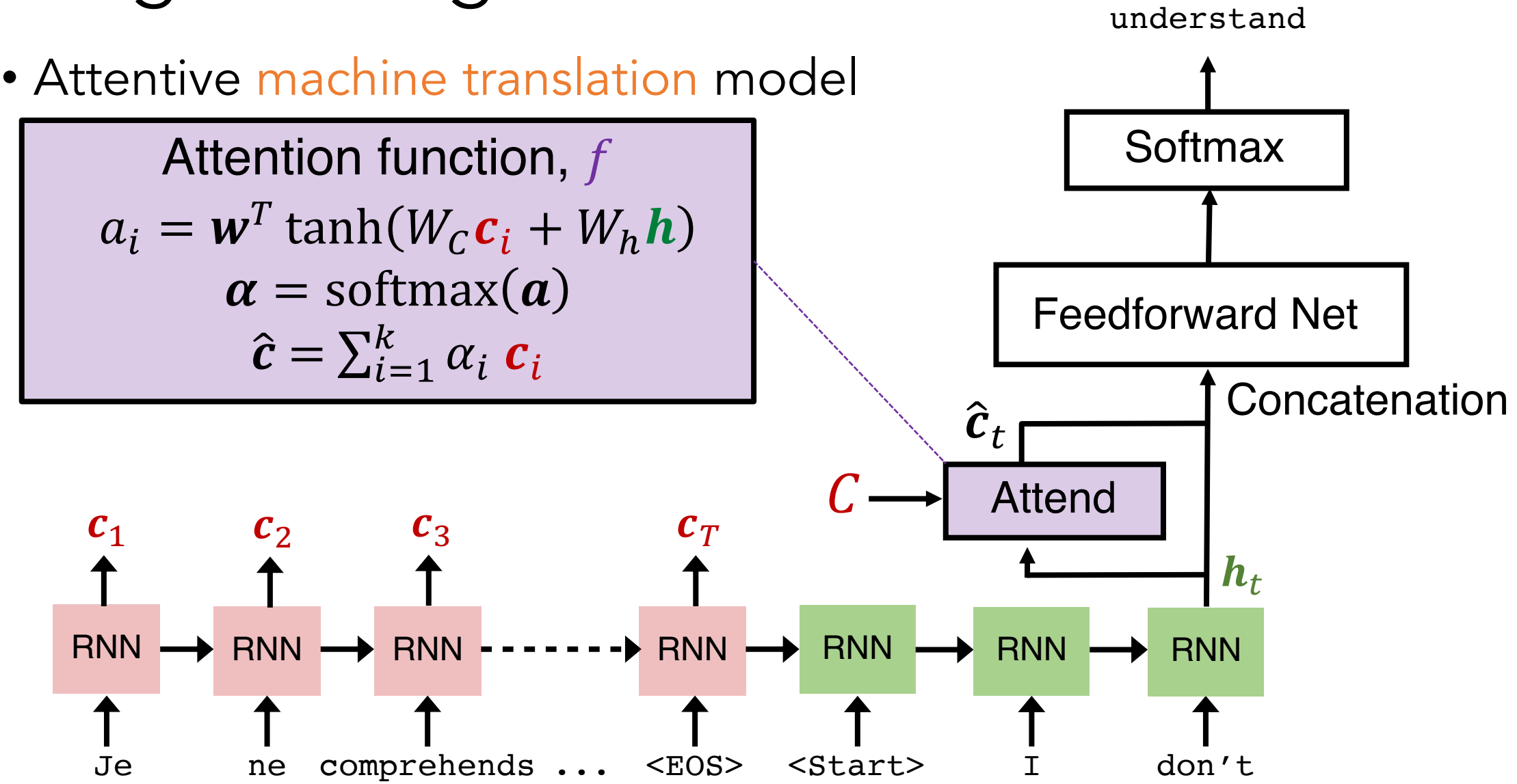

Attention output
me
$\hat{y}_3$
Attention distribution
Attention scores
Encoder RNN
Decoder RNN
il
a
m'
entarté
<START>
he
hit
Source sentence (input)



Slides Credit: Abigail See

Putting it all together

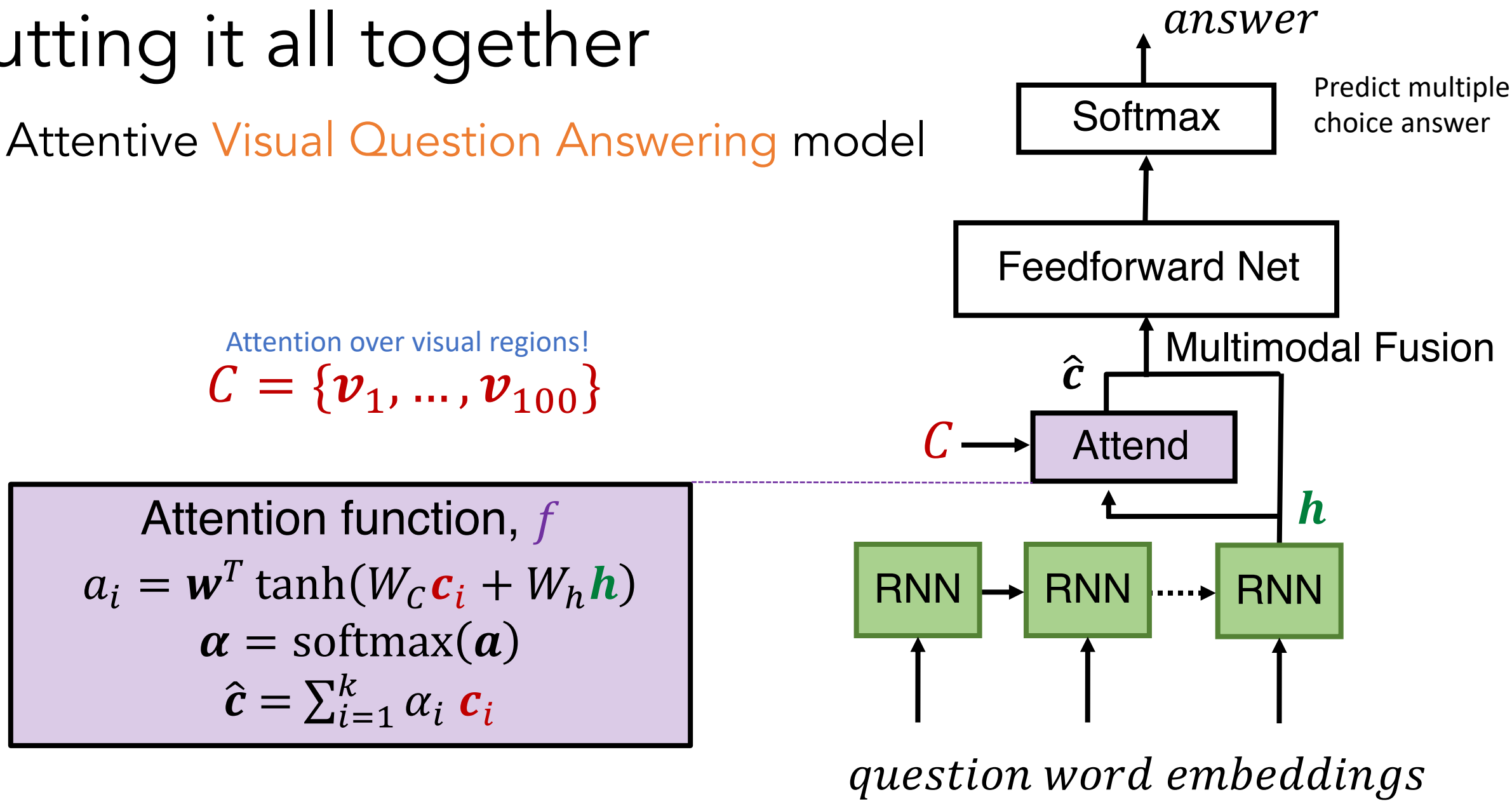
- Attentive machine translation model



Slides Credit: Peter Anderson

Putting it all together

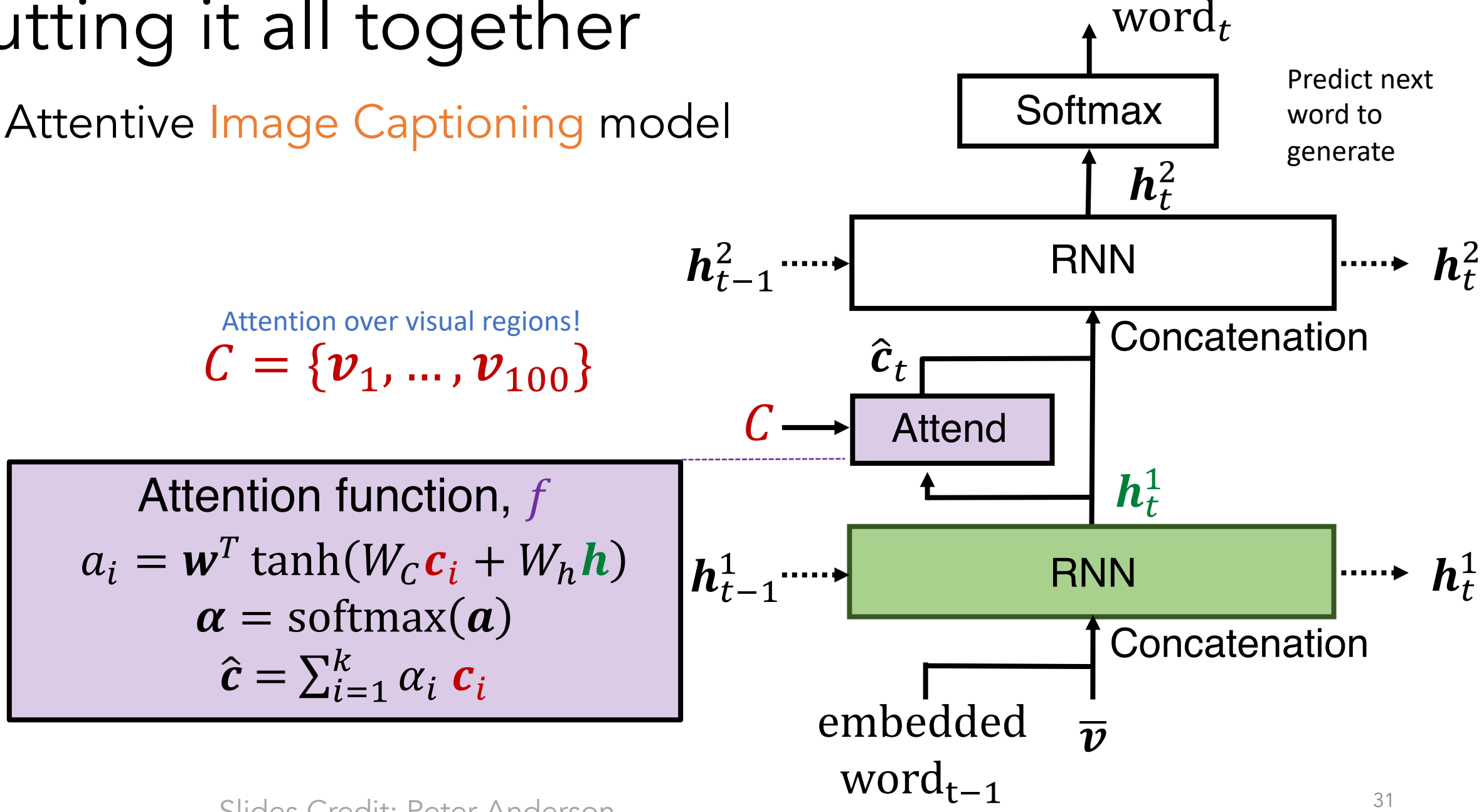
- Attentive Visual Question Answering model



Slides Credit: Peter Anderson

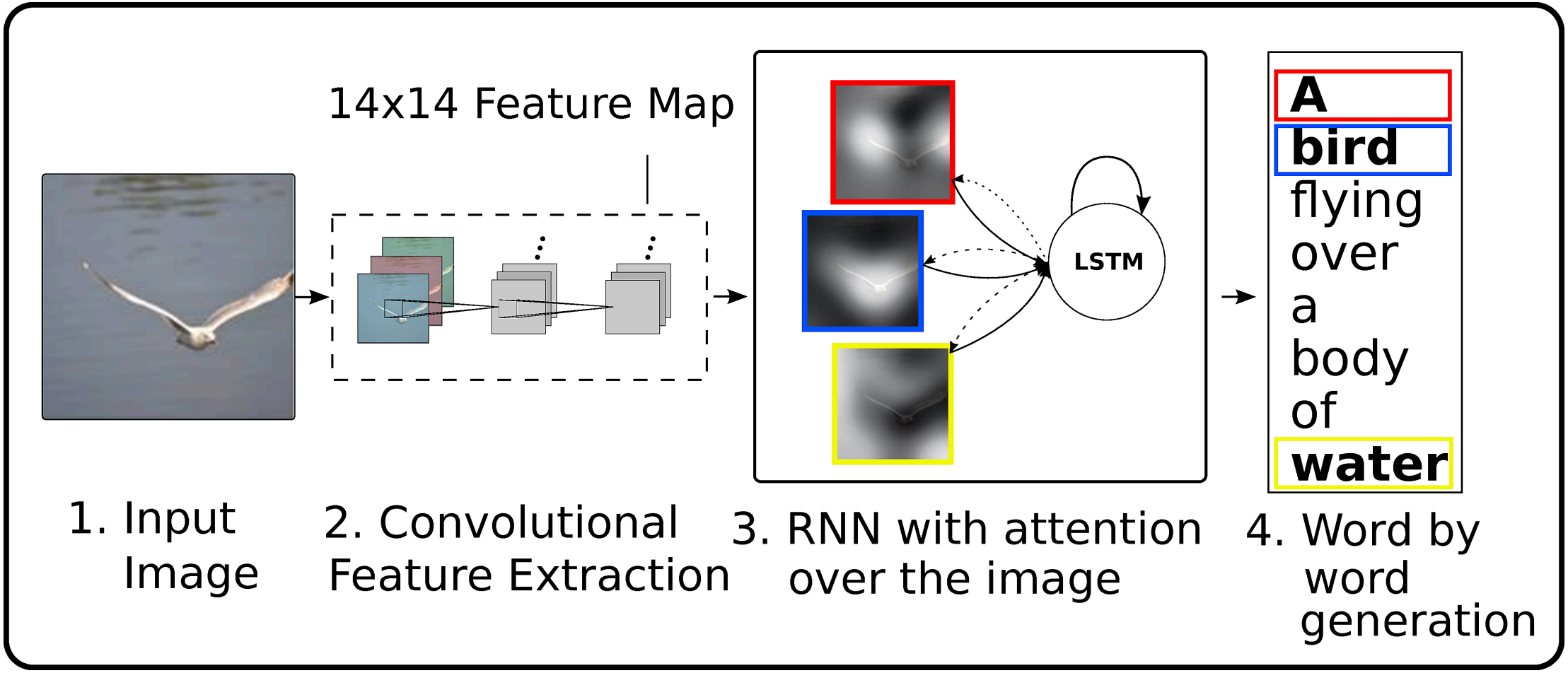
Putting it all together

- Attentive Image Captioning model



Slides Credit: Peter Anderson

Image captioning example

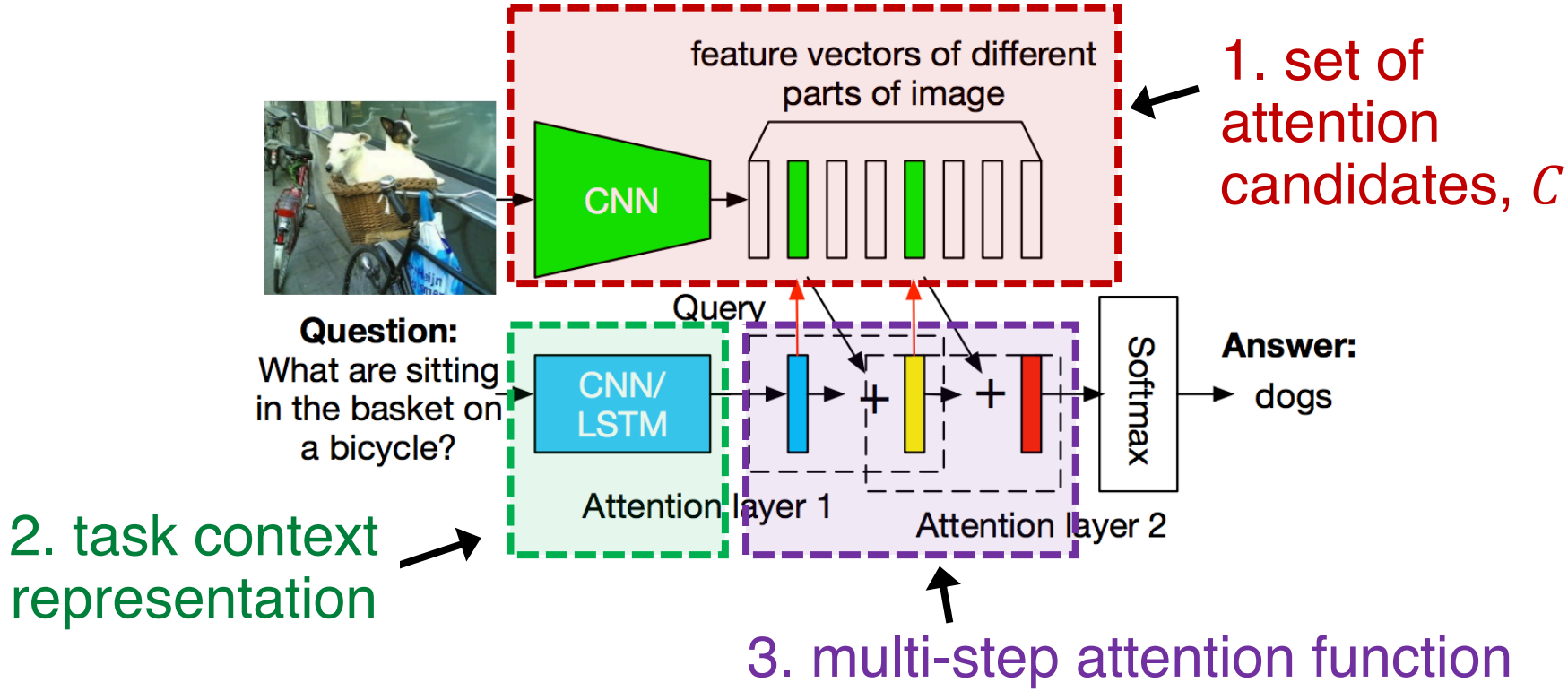


Xu et al. ICML 2015

Slides Credit: Peter Anderson

Stacked attention networks

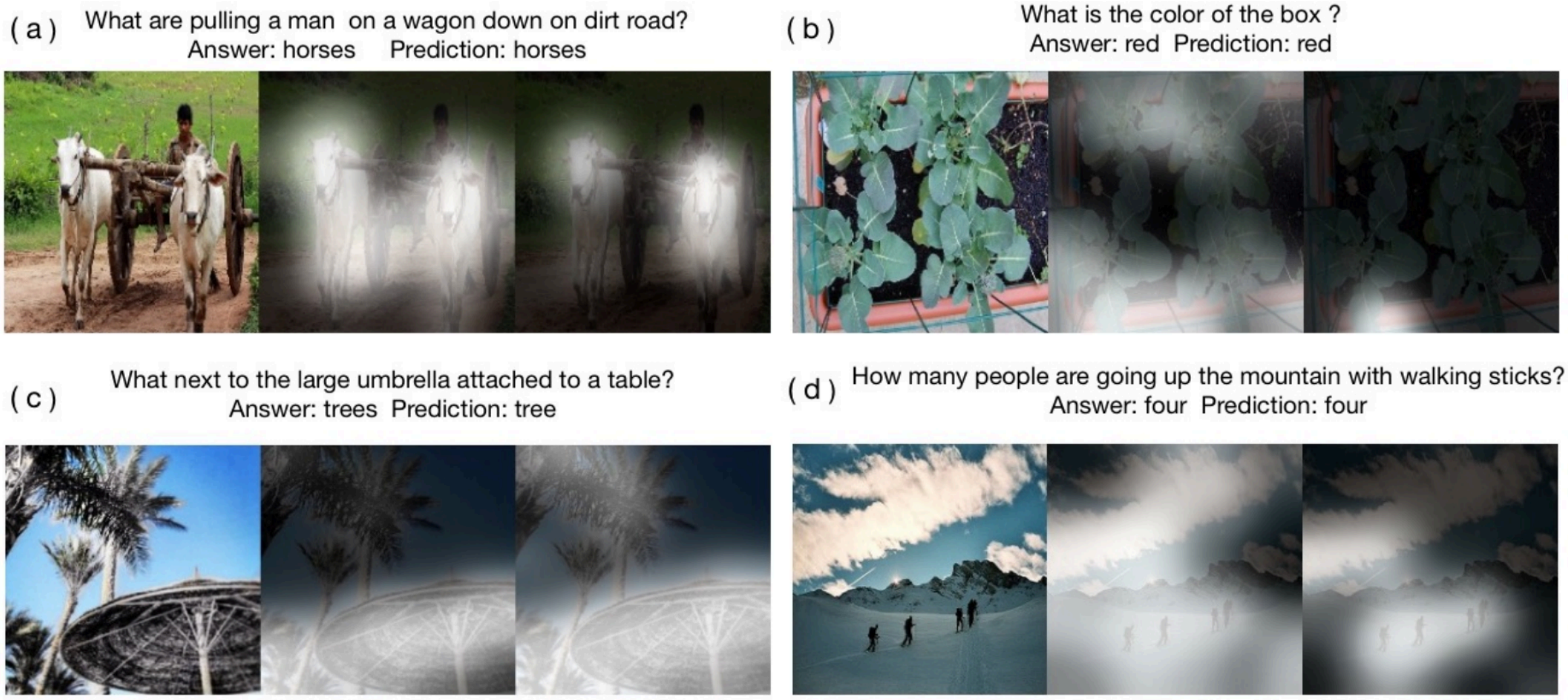
- Multiple attentions steps for VQA



Stacked Attention Networks for Image Question Answering, Yang et al. CVPR 2016

Slides Credit: Peter Anderson

Stacked attention networks

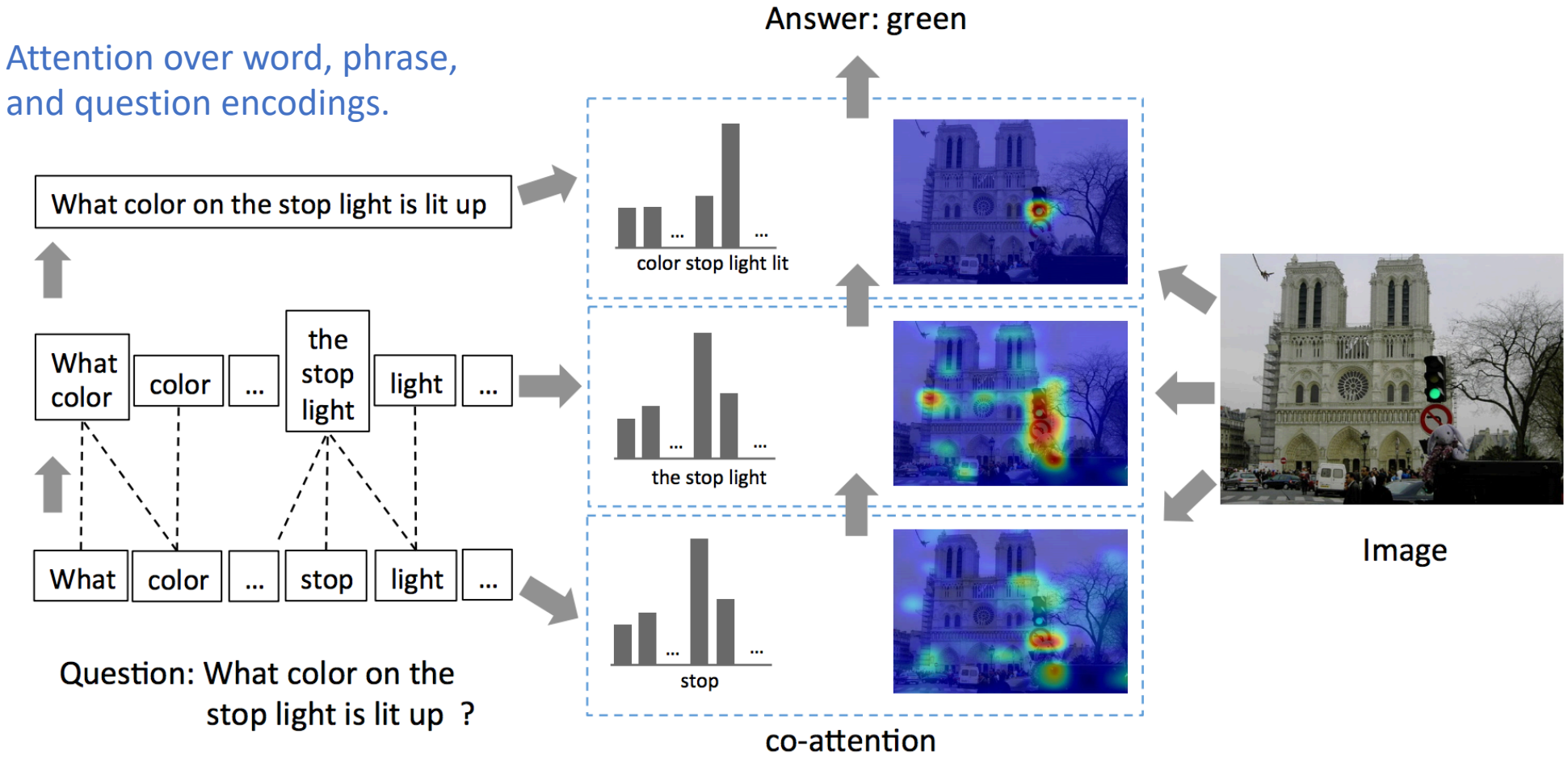


Stacked Attention Networks for Image Question Answering, Yang et al. CVPR 2016

Slides Credit: Peter Anderson

Hierarchical question-image co-attention

- Attending jointly to both question and image in VQA



Hierarchical Question-Image Co-Attention for Visual Question Answering, Lu et al. NIPS 2016

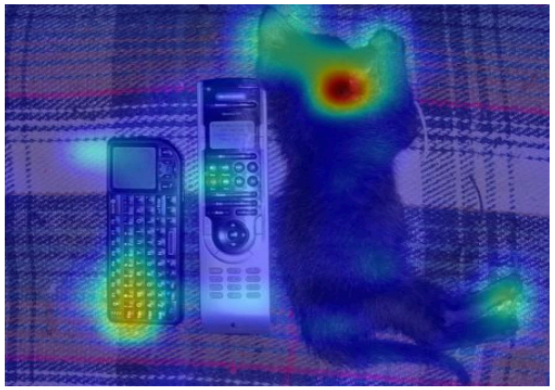
Slides Credit: Peter Anderson

Hierarchical question-image co-attention

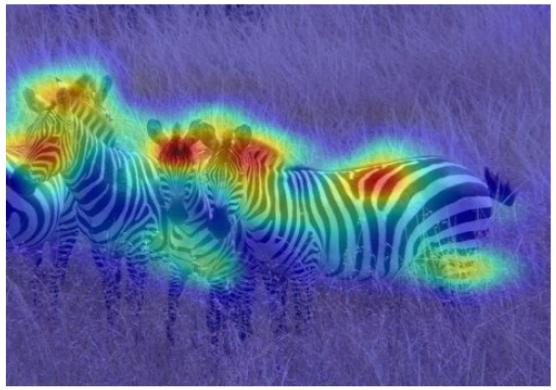
Q: what is the color of the kitten? **A**: **black**

Q: what are standing in tall dry grass look at the tourists? **A**: **zebras**

Q: where is the woman while her baby is sleeping? **A**: **kitchen**



what is the color of the kitten ?



what are standing in tall dry grass look at the tourists ?

where is the woman while her baby is sleeping ?

Hierarchical Question-Image Co-Attention for Visual Question Answering, Lu et al. NIPS 2016

Slides Credit: Peter Anderson

FiLM: Feature-wise Linear Modulation

- Applying attention by scaling and biasing CNN layers

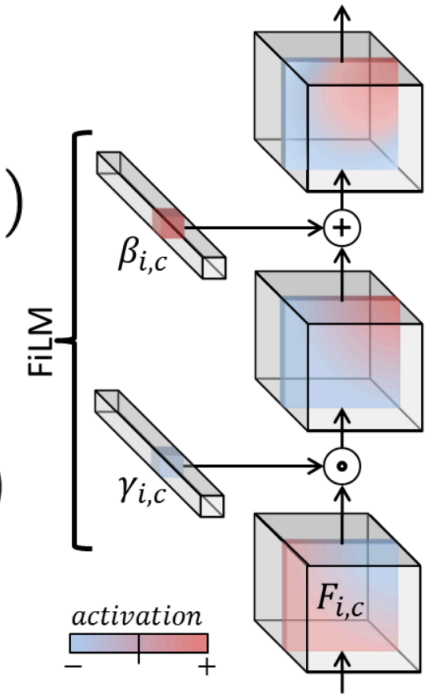
Feature-wise transformations are learned functions of input

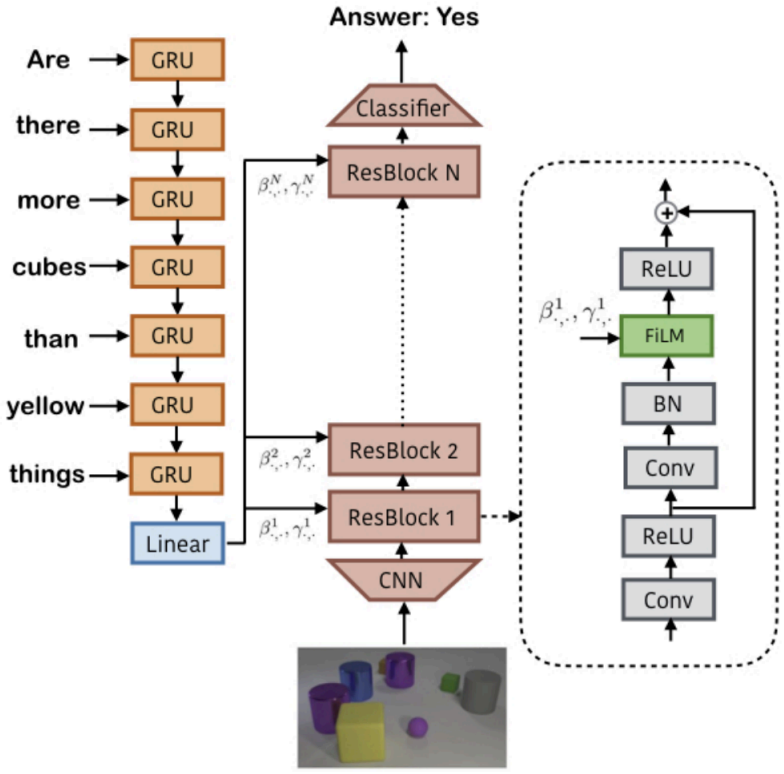
biasing

$$\beta_{i,c} = h_c(\boldsymbol{x}_i)$$

scaling

$$\gamma_{i,c} = f_c(\boldsymbol{x}_i)$$

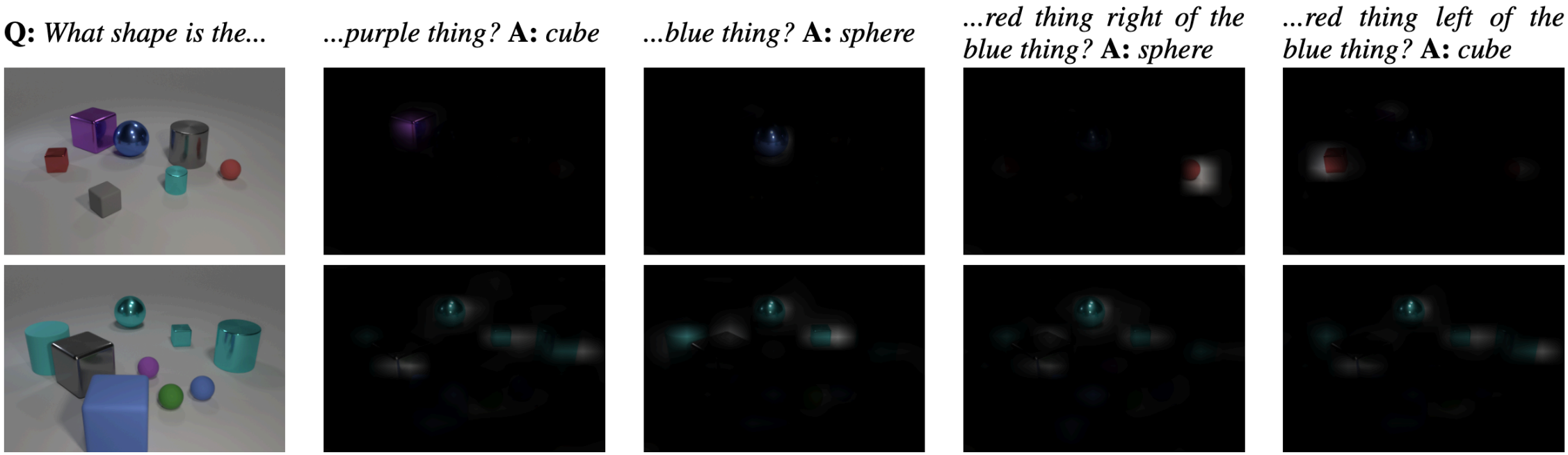




FiLM: Visual Reasoning with a General Conditioning Layer, Perez et al. AAAI 2018

Slides Credit: Peter Anderson

FiLM: Feature-wise Linear Modulation



FiLM: Visual Reasoning with a General Conditioning Layer, Perez et al. AAAI 2018

‘Soft’ vs ‘hard’ attention

- Soft: Each attention candidate is weighted by α_i

$$\hat{\boldsymbol{v}} = \sum_{i=1}^{k} \alpha_i \, \boldsymbol{v}_i$$

  - Easy to train (smooth and differentiable)
  - But can be expensive over large input

- Hard: Use α_i as a sample probability to pick *one* attention candidate as input to subsequent layers
  - Trainable with REINFORCE approaches (Xu et al. ICML 2015), or Gumbel-Softmax (Jang et al. ICLR 2017)

Soft

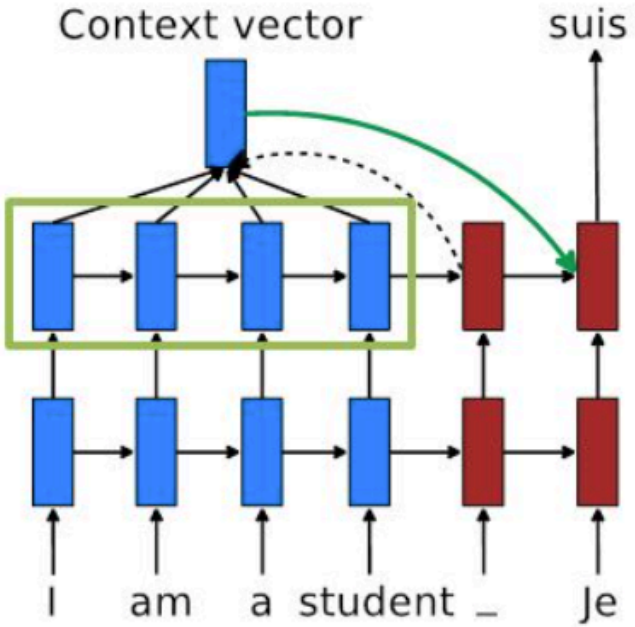
Hard

bird

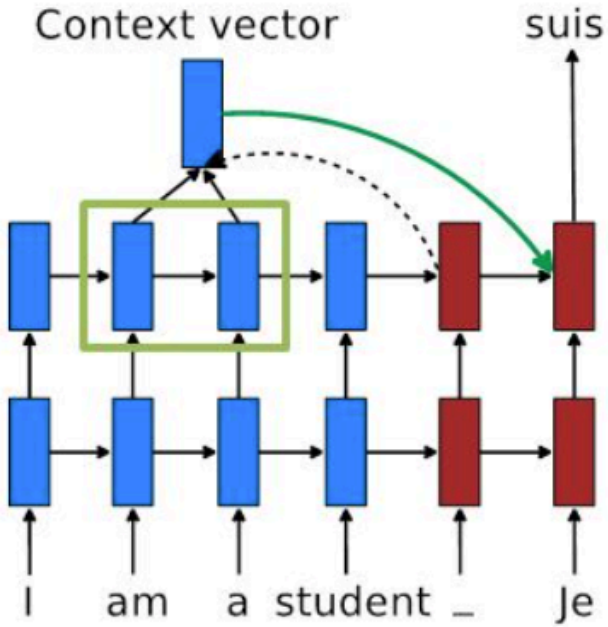
Xu et al. ICML 2015

‘Global’ vs ‘local’ attention

- Global: attention over the entire input
- Local: attention over a window (or subset) of the input



Global: ***all*** source states.



Local: ***subset*** of source states.

Luong et al, 2015

Attention

Attention According to Cognitive Psychology / Neuroscience

- A set of mechanisms that limit some processing to a subset of incoming stimuli (reducing computational demands)
- Can be driven 'top-down' by task demands (i.e. volitionally)
- Can be driven 'bottom-up' by salient stimuli (i.e. involuntarily)
- Visual attention can be applied to features, objects and spatial regions, as well as temporal cues (anticipating events)

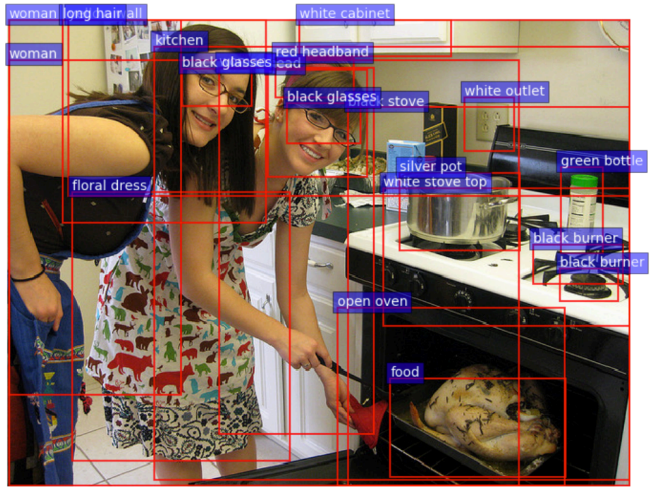
Buschman and Miller 2007, Scholl 2001

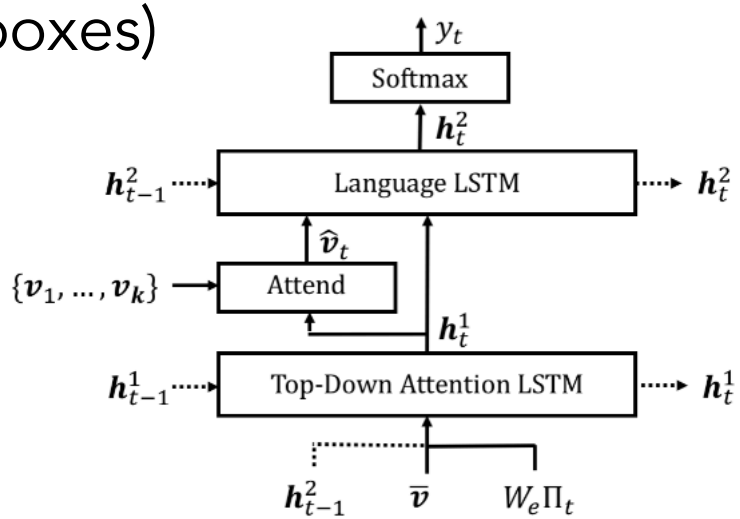
Slides Credit: Peter Anderson

Bottom-up and top-down attention

Combines 'top-down' and 'bottom-up' attention

- 'top-down' attention = *soft* attention over image conditioned on the task
  - VQA: question
  - Image captioning: what has been output before
- 'bottom-up' attention = *hard* attention using Faster-RCNN to identify image regions (object bounding boxes)





Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, Anderson et al. CVPR 2018

Slides Credit: Peter Anderson

Bottom-up and top-down attention

ResNet (10×10): A man sitting on a ~~toilet~~ in a bathroom.

Up-Down: A man sitting on a couch in a bathroom.

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, Anderson et al. CVPR 2018

Slides Credit: Peter Anderson

Bottom-up and top-down attention

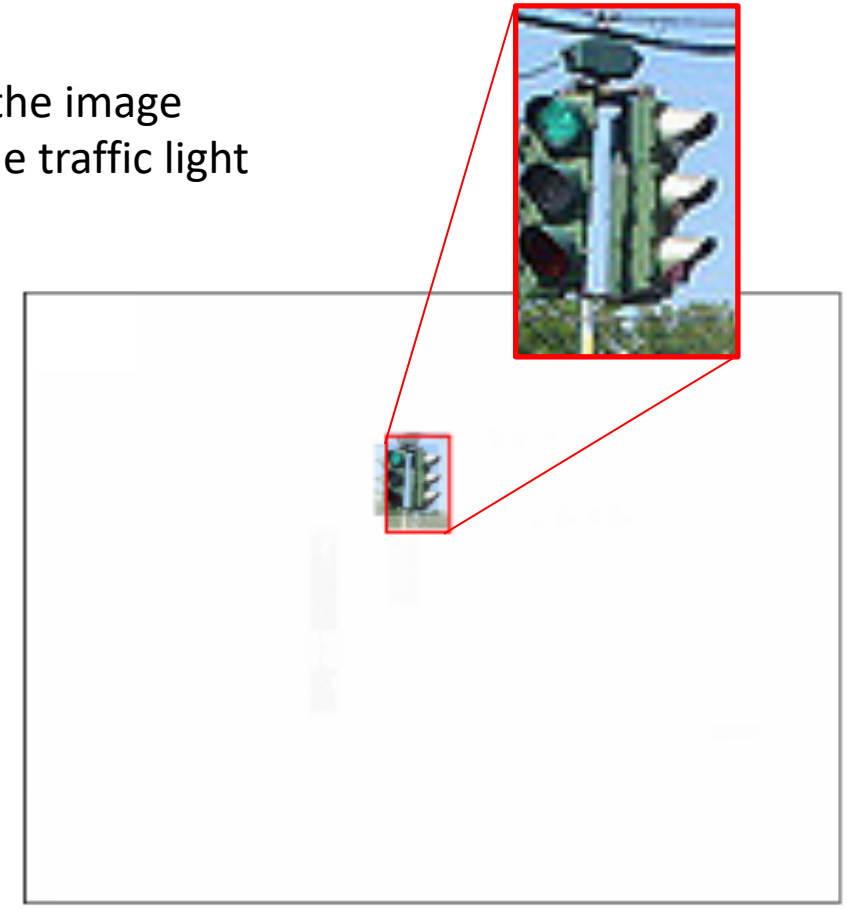
Q: What color is illuminated on the traffic light?

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, Anderson et al. CVPR 2018

Slides Credit: Peter Anderson

Bottom-up and top-down attention

Focus on region of the image corresponding to the traffic light



Q: What color is illuminated on the traffic light?

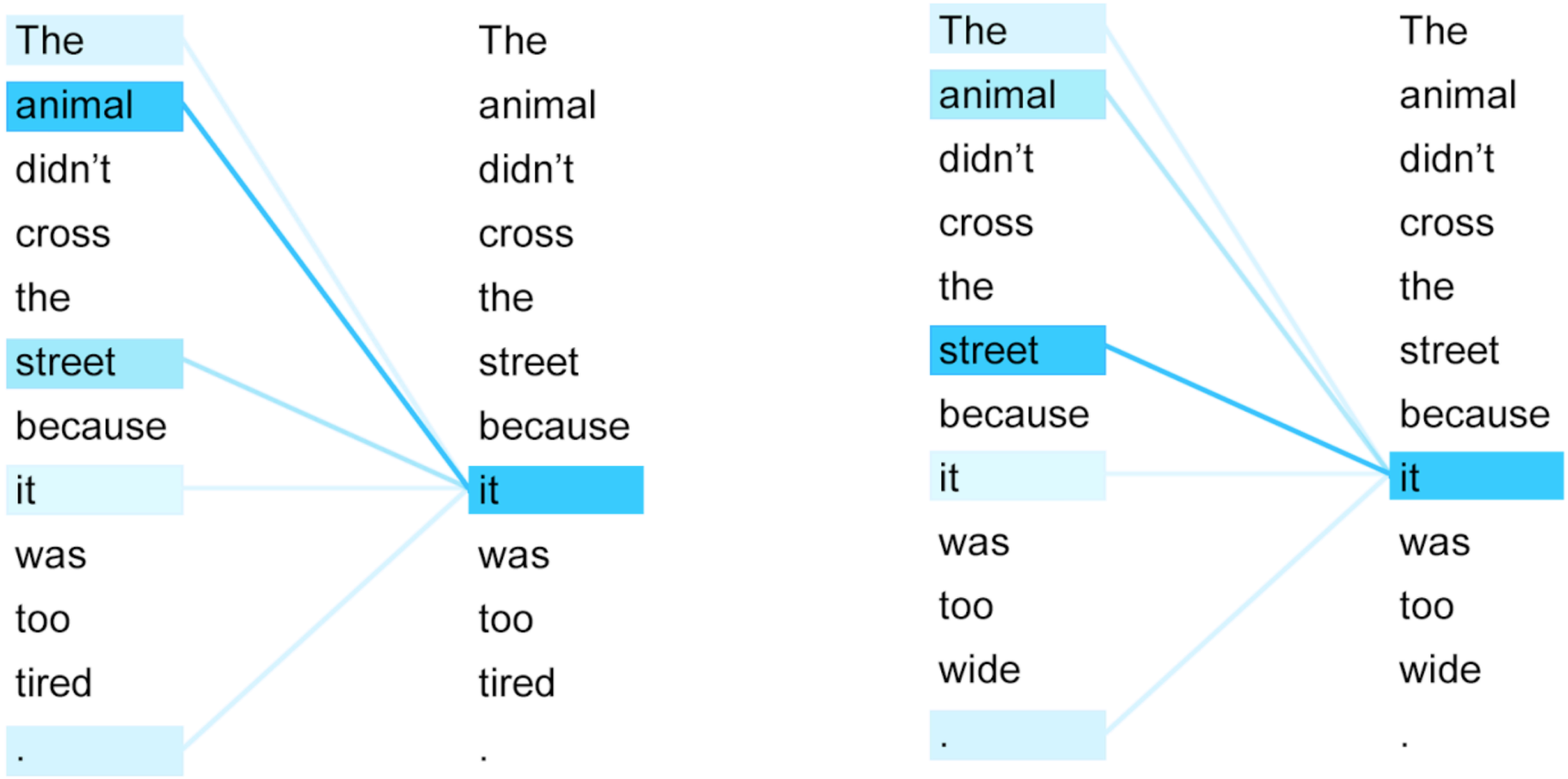
A: **green**

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, Anderson et al. CVPR 2018

Slides Credit: Peter Anderson

Self-attention

- Attention (correlation) with different parts of itself



- Transformers: modules with scaled dot-product self-attention

Types of attention scores

Attention function, f

$a_i = g(\boldsymbol{c}_i, \boldsymbol{z})$

$\boldsymbol{\alpha} = \text{softmax}(\boldsymbol{a})$

$\hat{\boldsymbol{c}} = \sum_{i=1}^{k} \alpha_i \, \boldsymbol{c}_i$

- Dot-product attention:

  $g(c_i, z) = z^\top c_i$

- Scaled dot-product attention:

  $g(c_i, z) = z^\top c_i / \sqrt{d}$

- Bilinear / multiplicative attention:

  $g(c_i, z) = z^\top W c_i \in \mathbb{R}$

  where W is a weight matrix

- Additive attention (essentially MLP):

  $g(c_i, z) = v^\top \tanh\left(W_1 c_i + W_2 z\right)$

  where W_1, W_2 are weight matrices and v is a weight vector

Query-key-value view of attention

Attention function, f

$$a_i = g(\boldsymbol{c}_i, \boldsymbol{z})$$
$$\boldsymbol{\alpha} = \text{softmax}(\boldsymbol{a})$$
$$\hat{\boldsymbol{c}} = \sum_{i=1}^{k} \alpha_i \, \boldsymbol{c}_i$$

→

Attention function, f

$$a_i = g(\boldsymbol{k}_i, \boldsymbol{q})$$
$$\boldsymbol{\alpha} = \text{softmax}(\boldsymbol{a})$$
$$\hat{\boldsymbol{c}} = \sum_{i=1}^{k} \alpha_i \, \boldsymbol{v}_i$$

Projected query,key,value →

$$\boldsymbol{q} = W_Q \, \boldsymbol{z}$$
$$\boldsymbol{k}_i = W_K \, \boldsymbol{c}_i$$
$$\boldsymbol{v}_i = W_V \, \boldsymbol{c}_i$$

→ Matrix form

$$\boldsymbol{q} = W_Q \, \boldsymbol{z}$$
$$K = W_K \, C^T$$
$$V = W_V \, C^T$$

More next Thursday when we discuss transformers

Next week

- Monday: Paper presentations and discussions
  - Show, attend, and tell (captioning)
    - Guest presenter: Ali Gholami
  - MattNet (referring expressions)
- Thursday: Pretraining with Transformers