CS839:
Probabilistic Graphical Models

Lecture 22: The Attention Mechanism

Theo Rekatsinas

Why Attention?

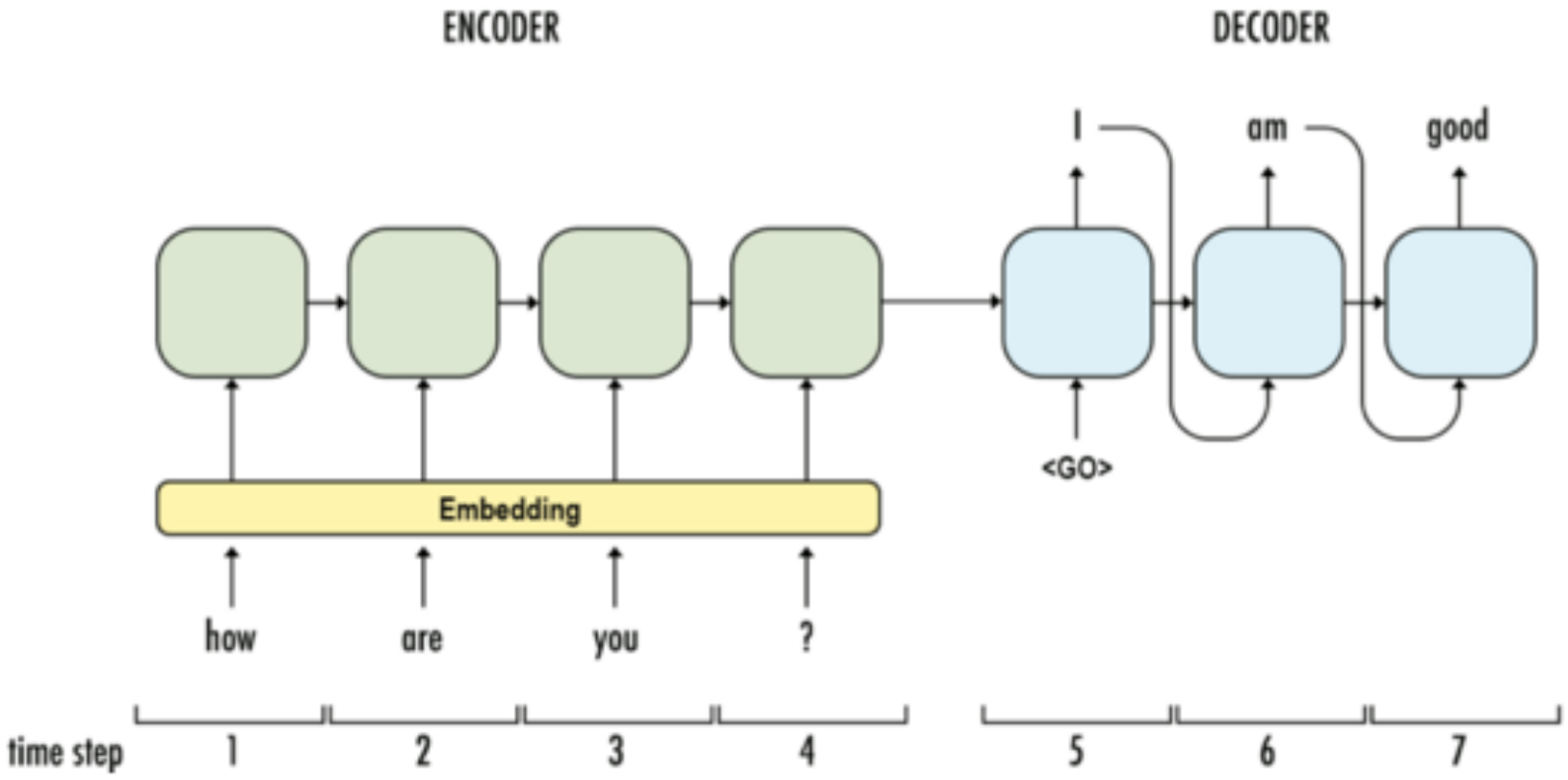
- Consider machine translation:
  - We need to pay attention to the word we are currently translating. Is the entire sequence needed as context?

  - The cat is black -> Le chat est noir

Why Attention?

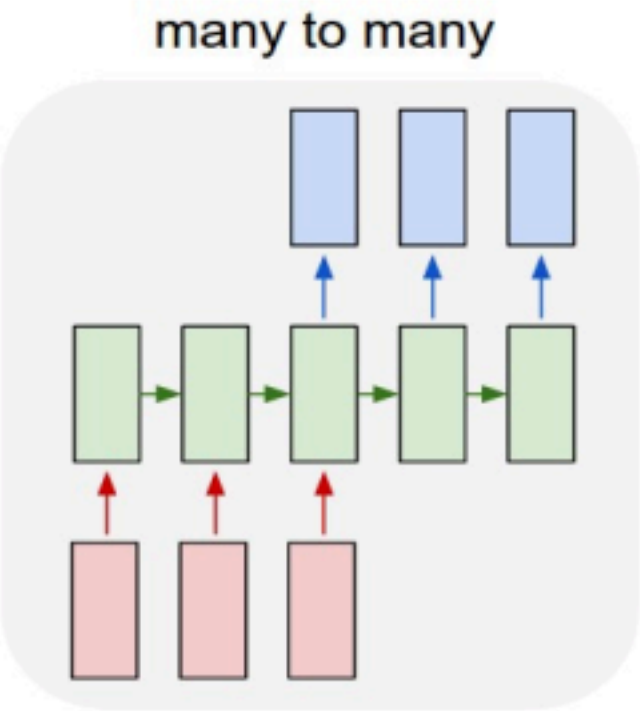
- Consider machine translation:
  - We need to pay attention to the word we are currently translating. Is the entire sequence needed as context?
  - The cat is black -> Le chat est noir
- RNNs are the de-facto standard for machine translation
- Problem: translation relies on reading a complete sentence and compresses all information into a fixed-length vector a sentence with hundreds of words represented by several words will surely lead to information loss, inadequate translation, etc.
- Long-range dependencies are tricky.

Basic encoder - decoder



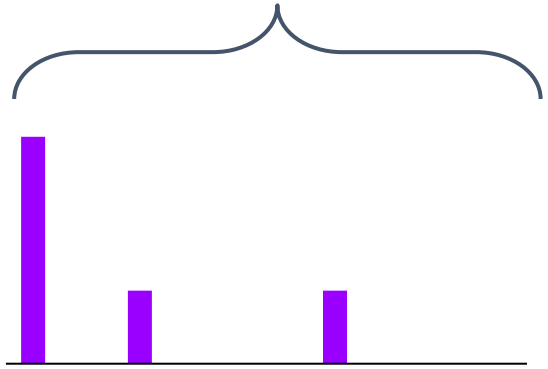
Soft Attention for Translation

“I love coffee” -> “Me gusta el café”

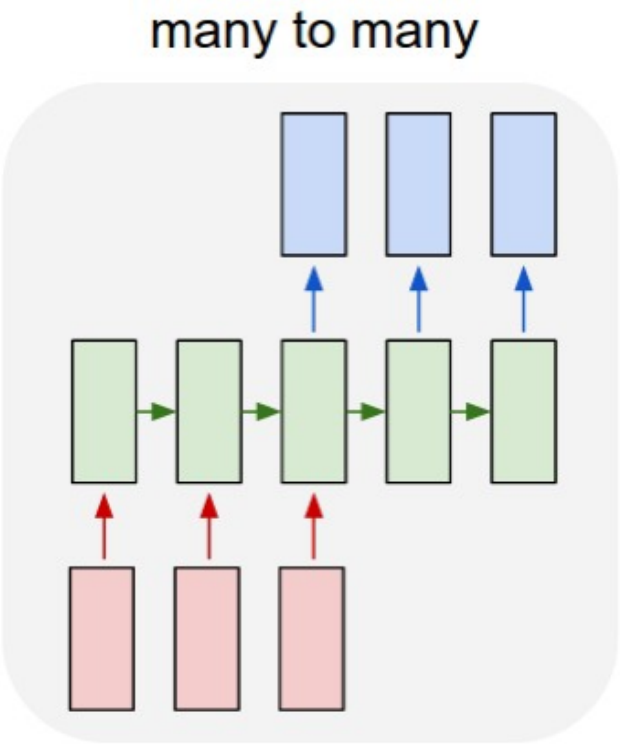


Soft Attention for Translation

Distribution over
input words



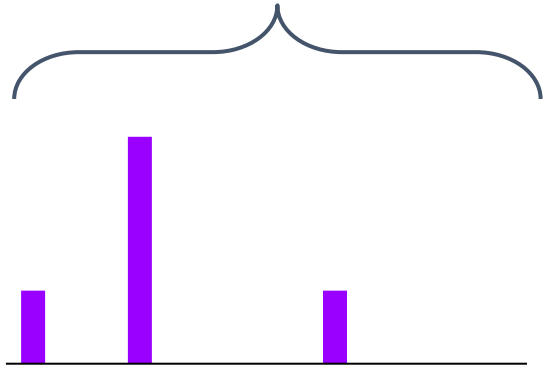
“I love coffee” -> “Me gusta el café”



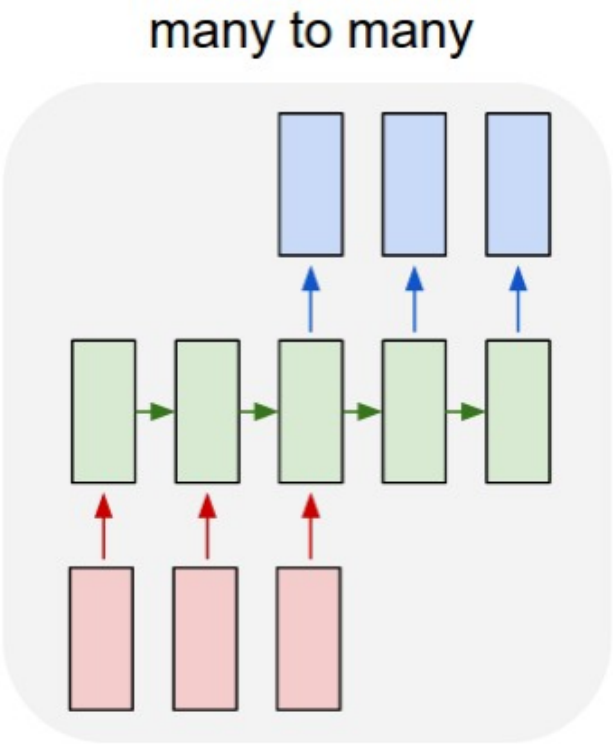
Bahdanau et al, “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015

Soft Attention for Translation

Distribution over input words



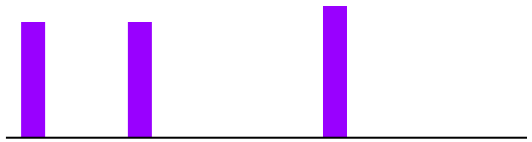
“I love coffee” -> “Me gusta el café”



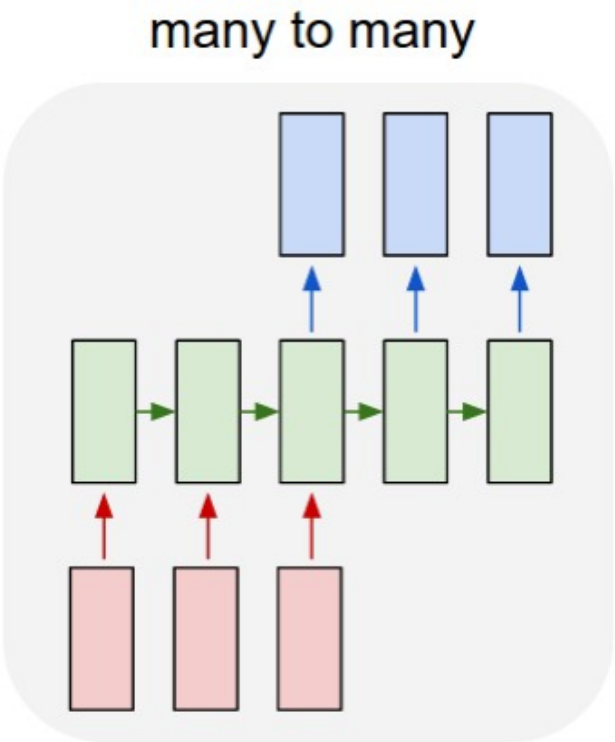
Bahdanau et al, “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015

Soft Attention for Translation

Distribution over
input words



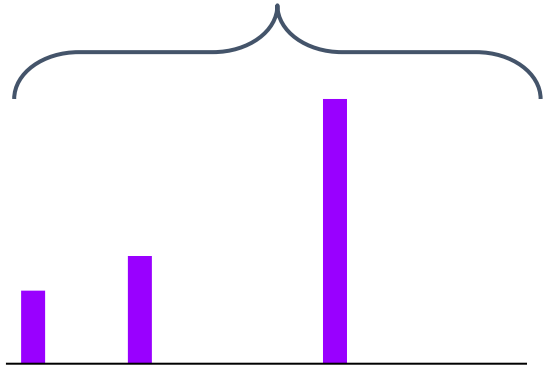
“I love coffee” -> “Me gusta el café”



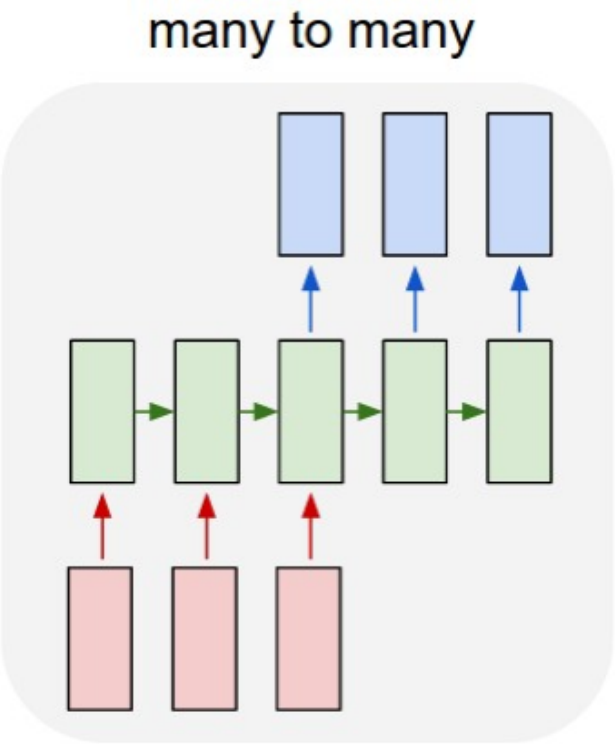
Bahdanau et al, “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015

Soft Attention for Translation

Distribution over input words

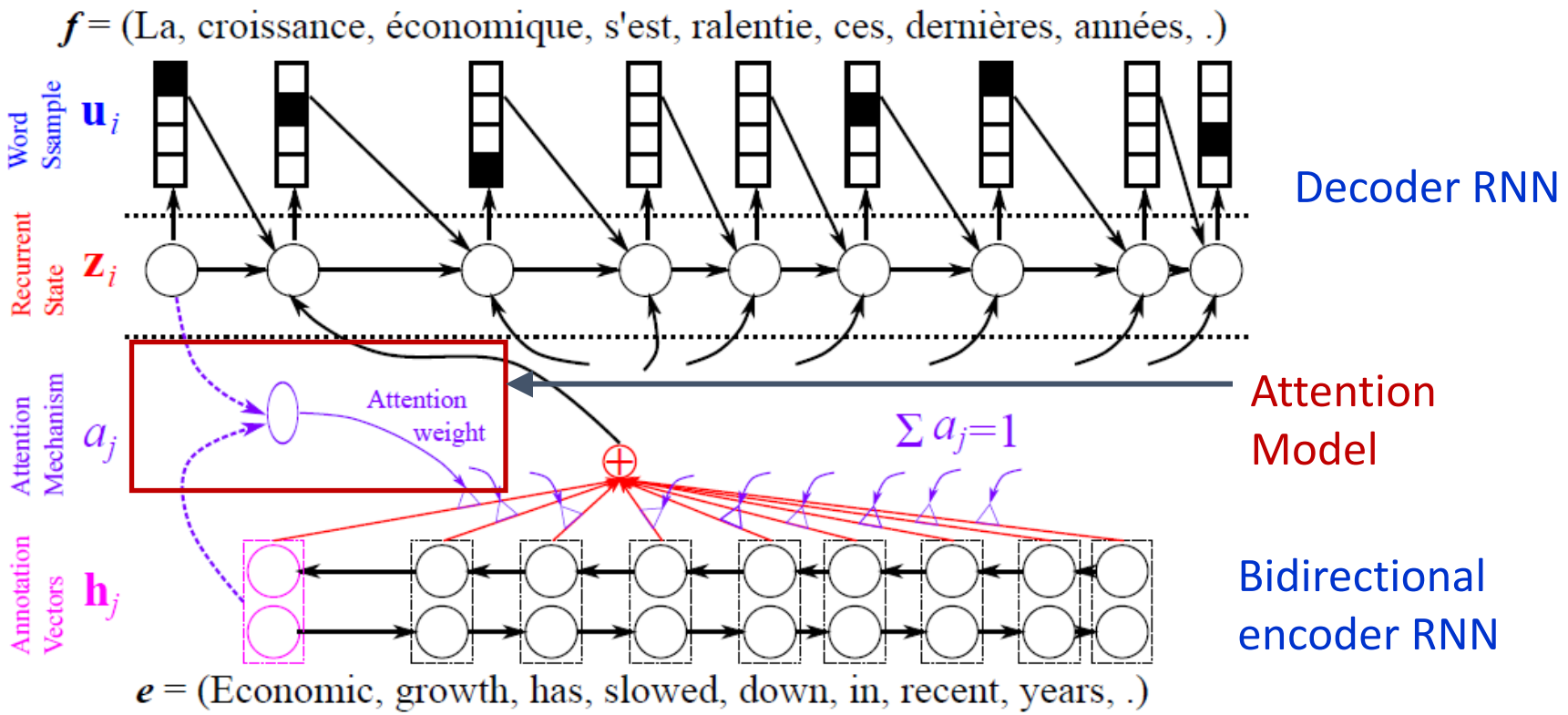


“I love coffee” -> “Me gusta el café”



Bahdanau et al, “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015

Soft Attention



From Y. Bengio CVPR 2015 Tutorial

Soft Attention

Context vector (input to decoder):

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

Mixture weights:

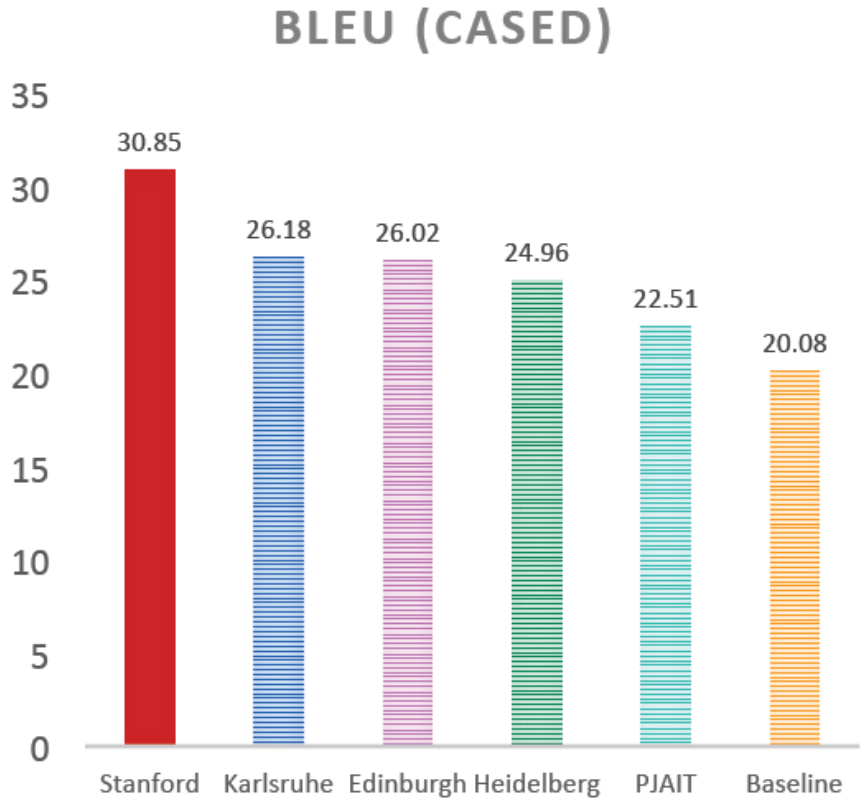
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

Alignment score (how well do input words near j match output words at position i):

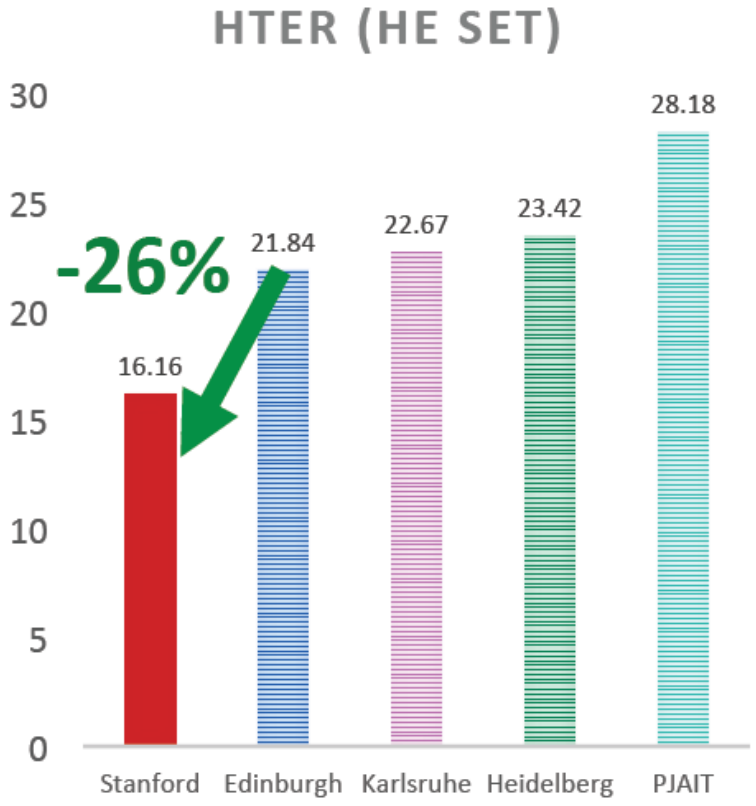
$$e_{ij} = a(s_{i-1}, h_j)$$

Soft Attention

Luong, Pham and Manning's Translation System (2015):



Translation Error Rate vs Human



Luong and Manning IWSLT 2015

Hard Attention

- Instead of a soft interpolation, make a **zero-one decision** about where to attend (Xu et al. 2015)
  - Harder to train, requires methods such as reinforcement learning (see later classes)
- Perhaps this helps interpretability? (Lei et al. 2016)

Review

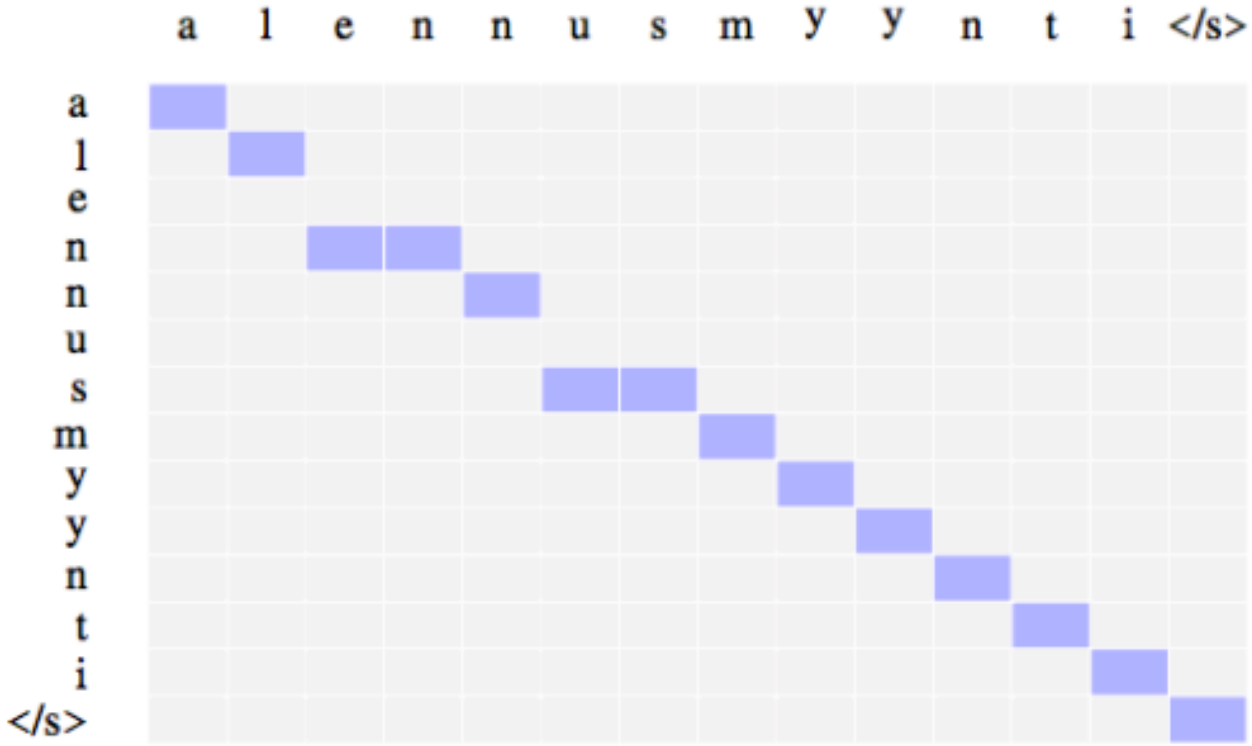
the beer was n't what i expected, and i'm not sure it's "true to style", but i thought it was delicious. **a very pleasant ruby red-amber color** with a relatively brilliant finish, but a limited amount of carbonation, from the look of it. aroma is what i think an amber ale should be - a nice blend of caramel and happiness bound together.

Ratings

Look: 5 stars *Smell:* 4 stars

Monotonic Attention

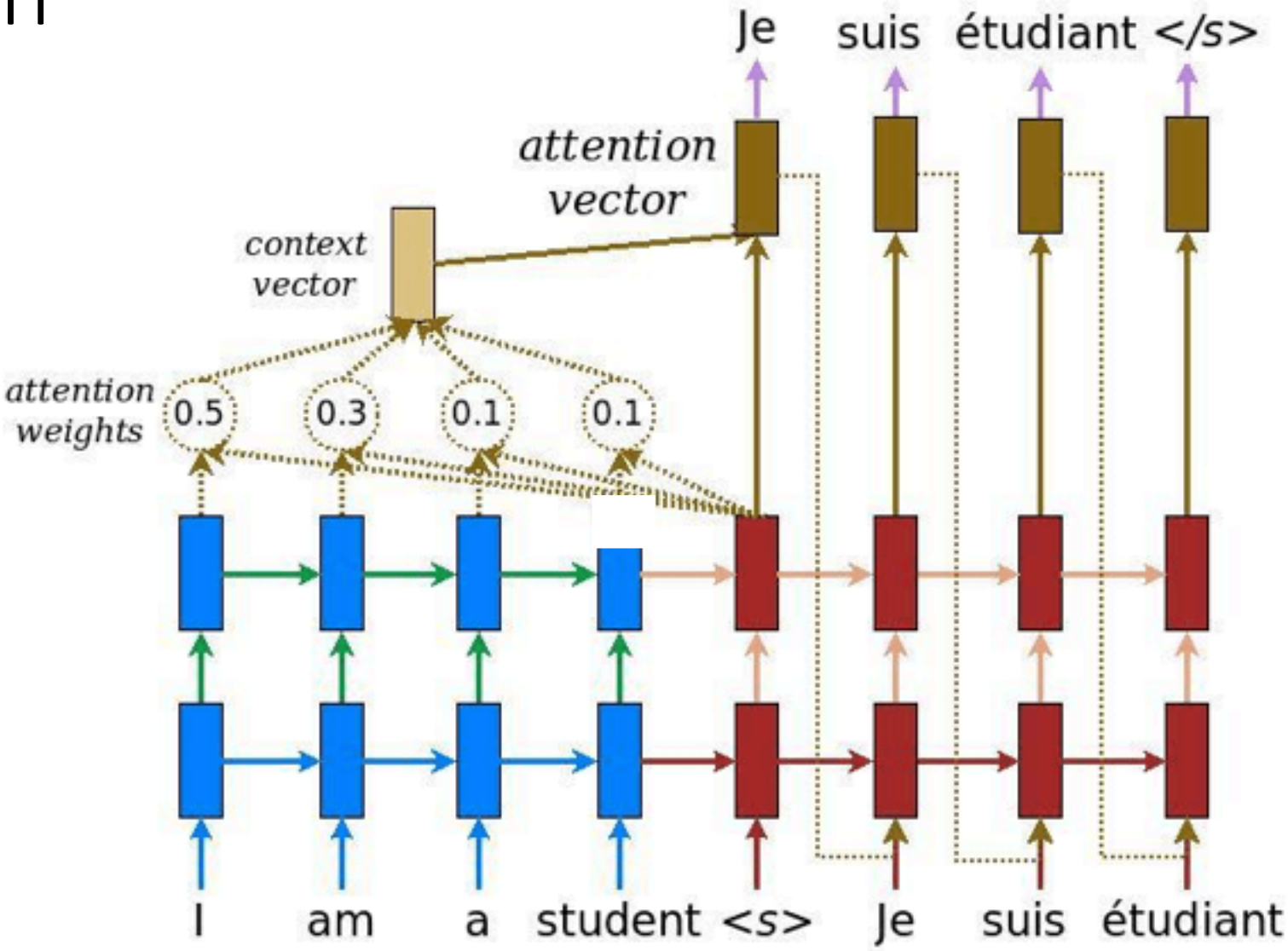
- In some cases, we might know the output will be the same order as the input
  - Speech recognition, incremental translation, morphological inflection (?), summarization (?)



- **Basic idea:** hard decisions about whether to read more

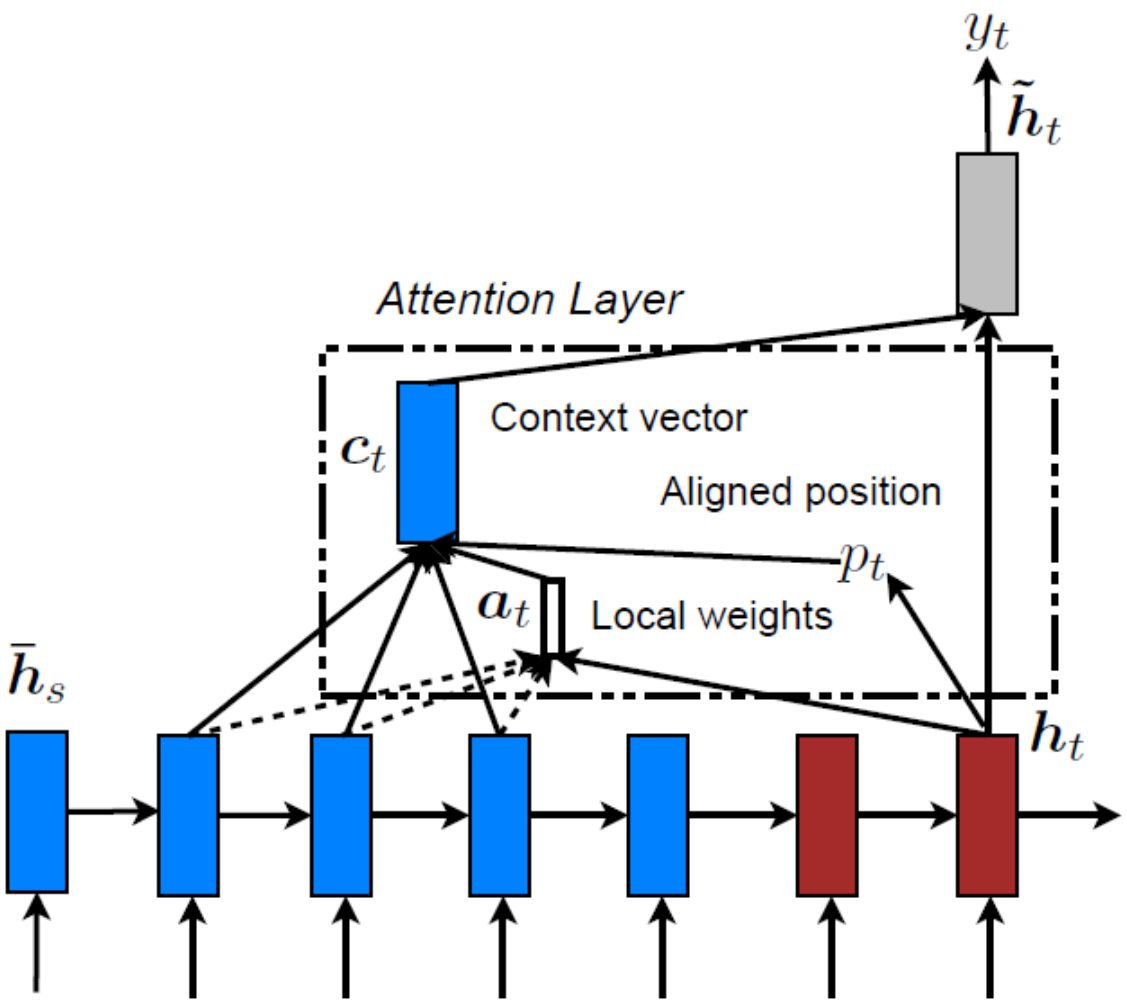
Global Attention

- Blue = encoder
- Red = decoder
- Attend to a context vector.
- Decoder captures global information not only the information from one hidden state.
- Context vector takes all cell's outputs as input and computes a probability distribution for each token the decoder wants to generate

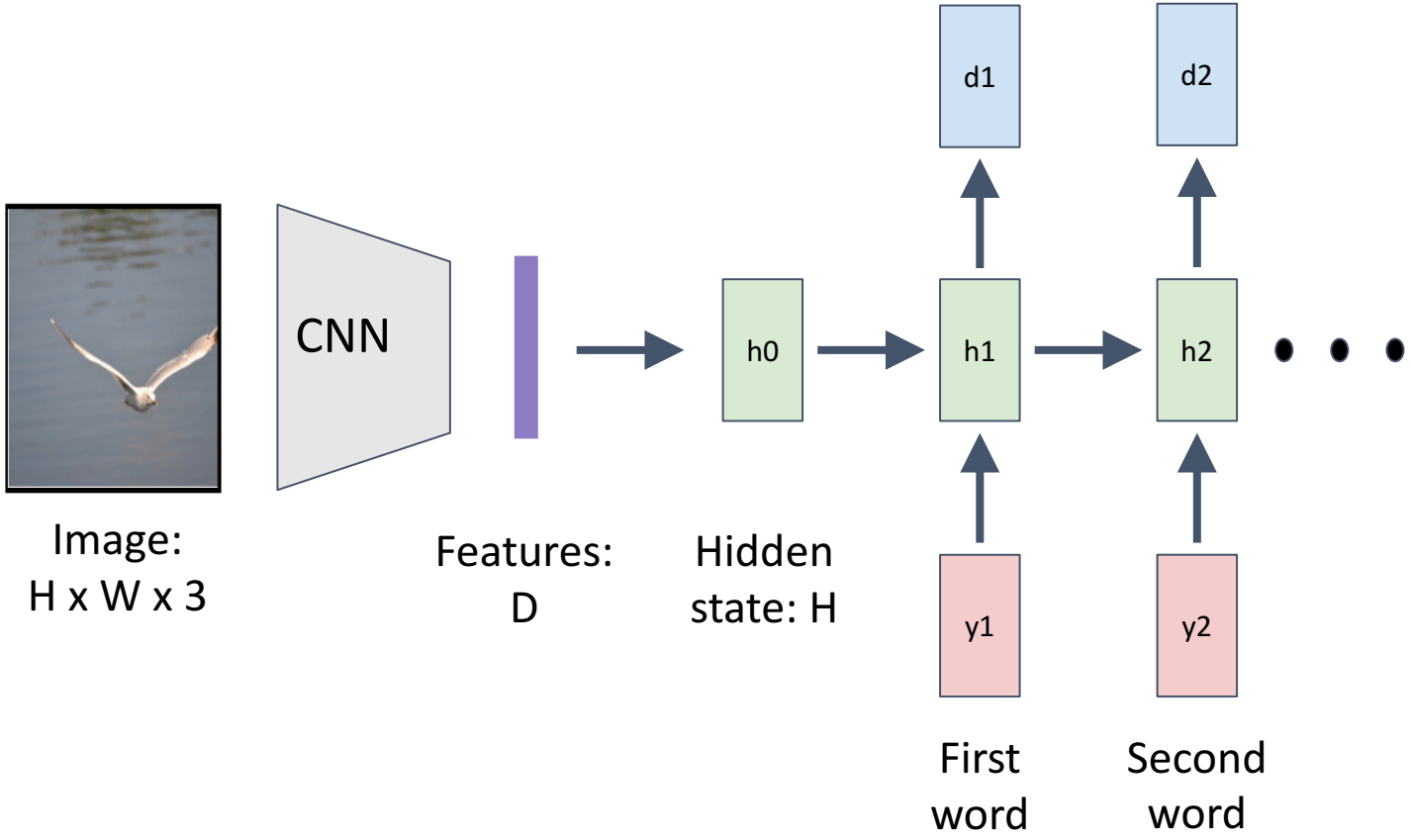


Local Attention

- Compute a best aligned position first
- Then compute a context vector centered at that position



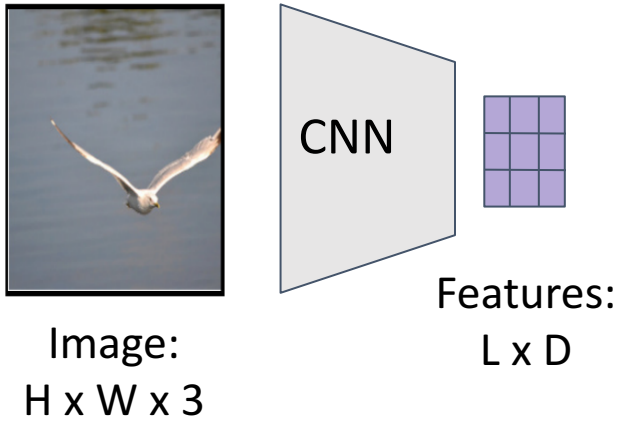
RNN for Captioning



RNN only looks at whole image, once

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

Soft Attention for Captioning



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

Soft Attention for Captioning

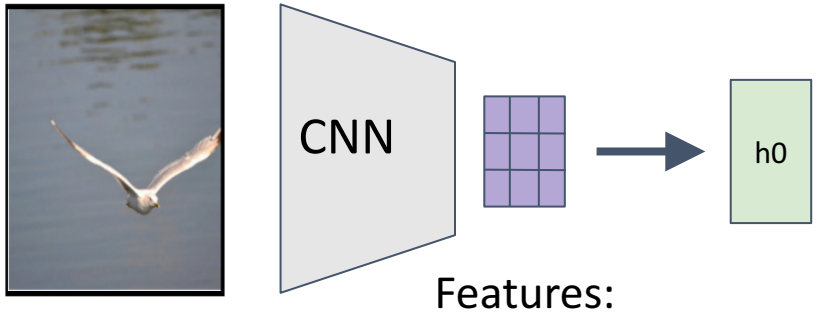
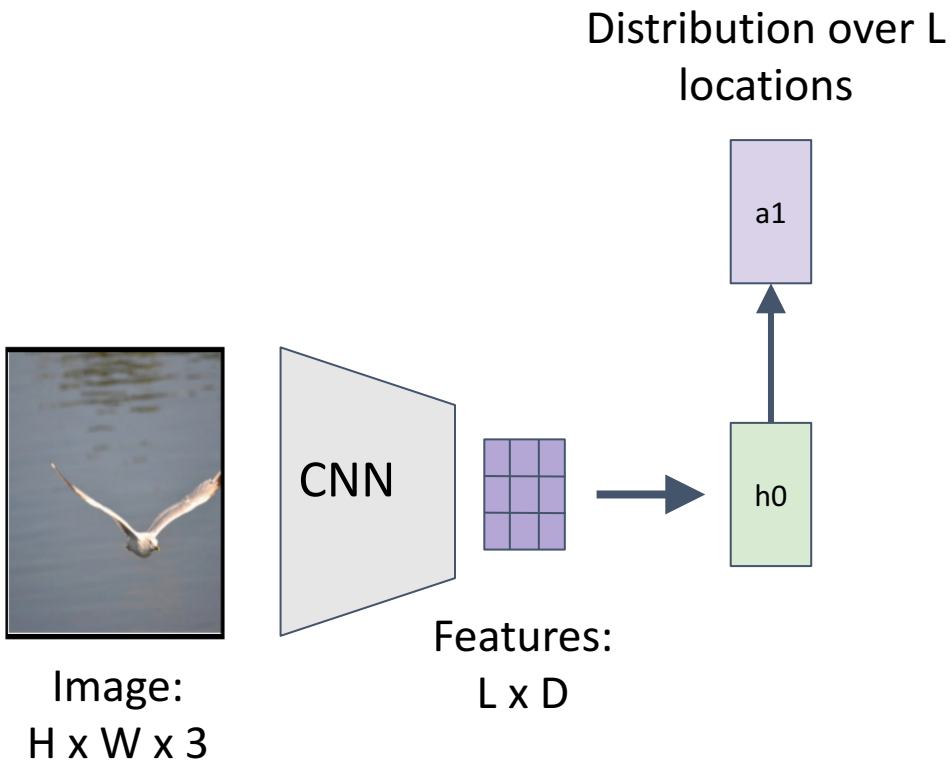


Image:
H x W x 3

Features:
L x D

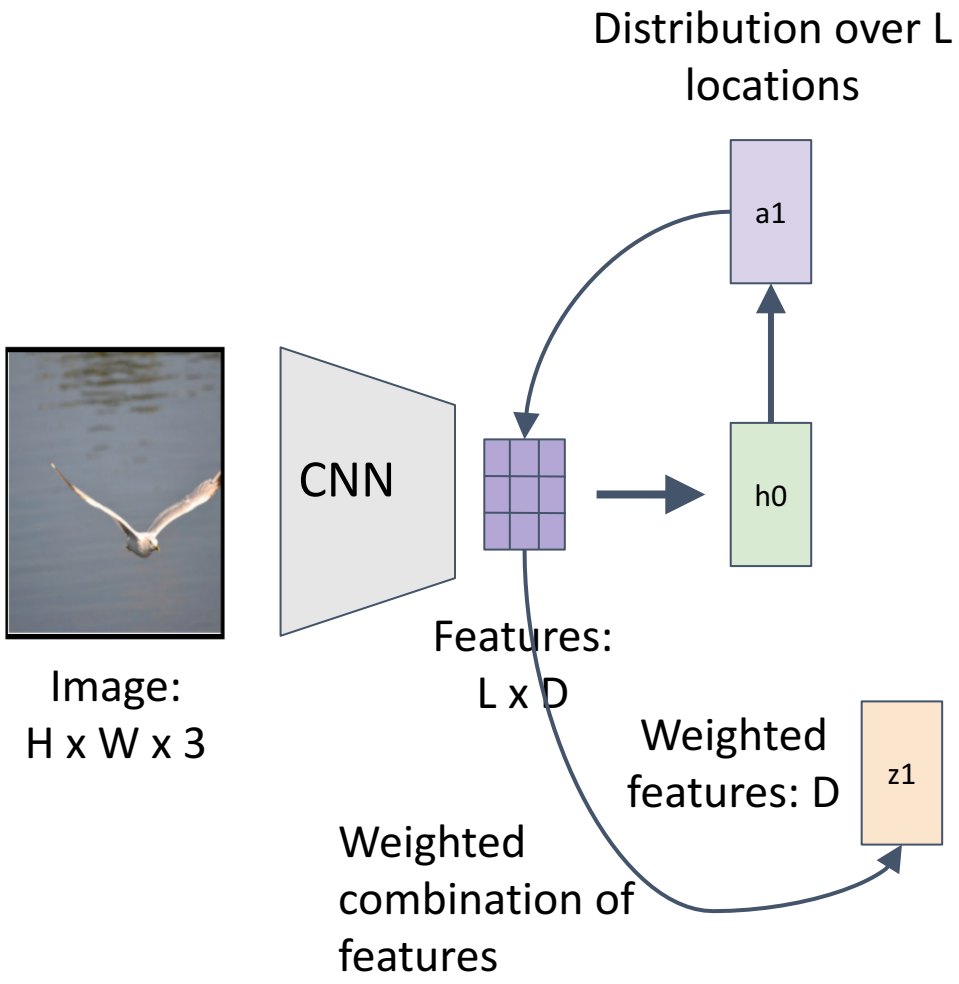
Xu et al, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML 2015

Soft Attention for Captioning

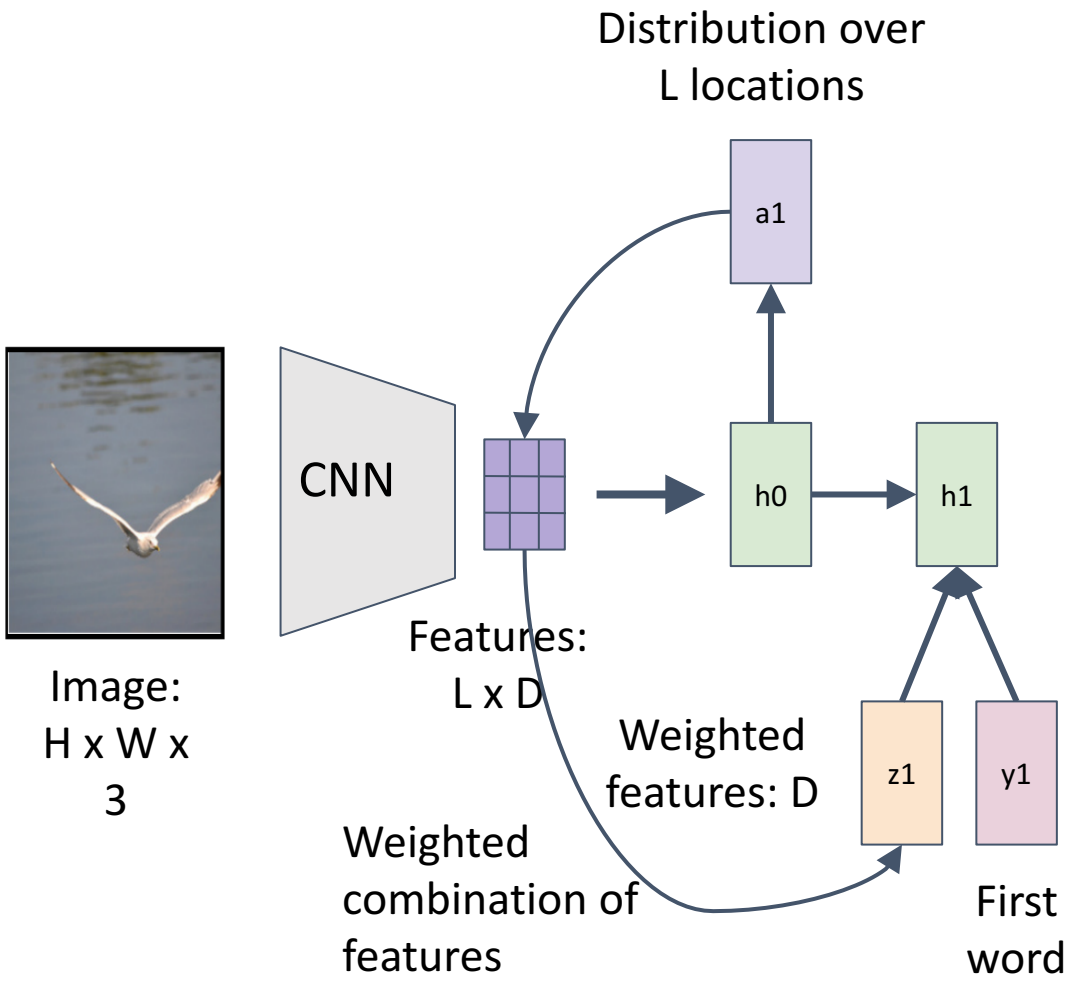


Xu et al, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML 2015

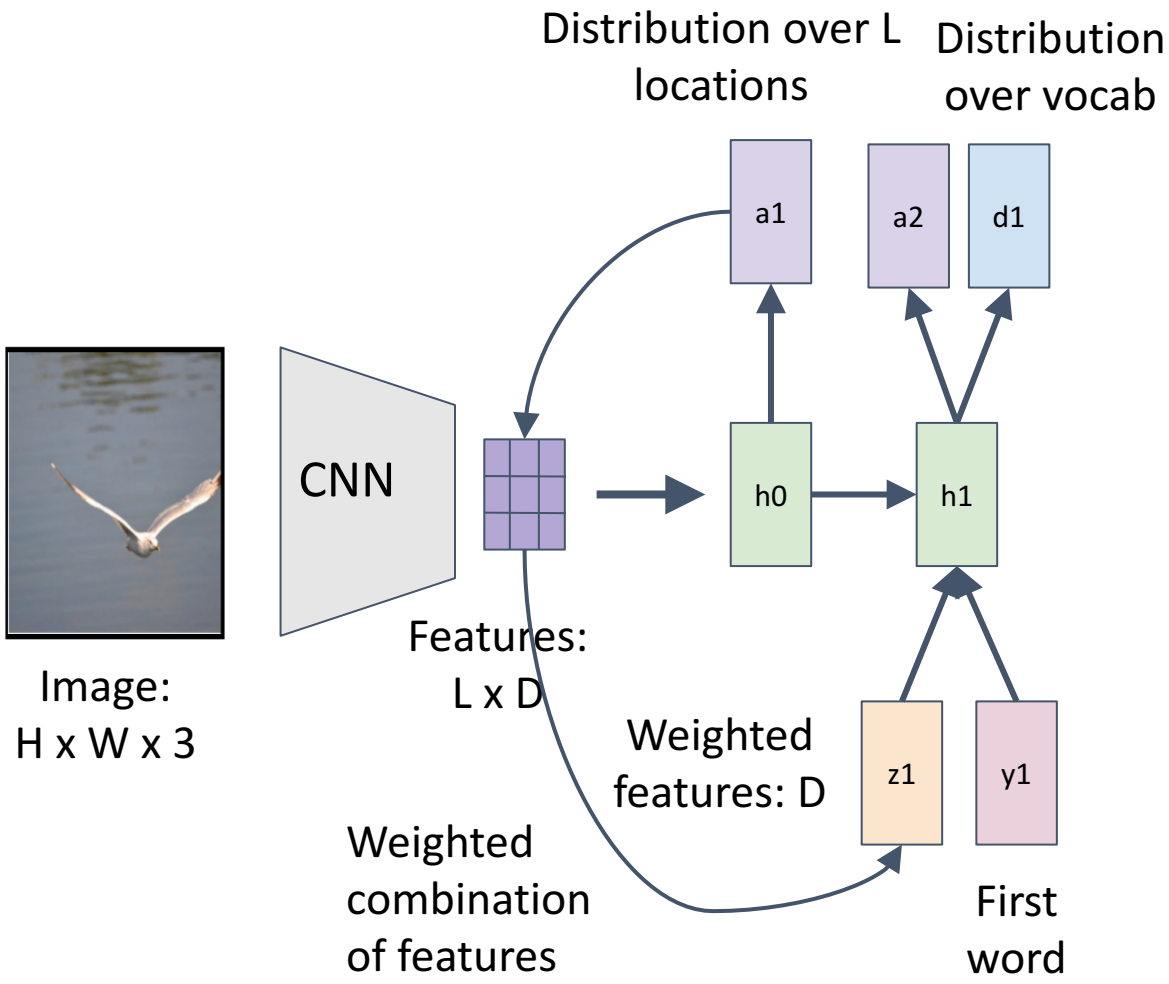
Soft Attention for Captioning



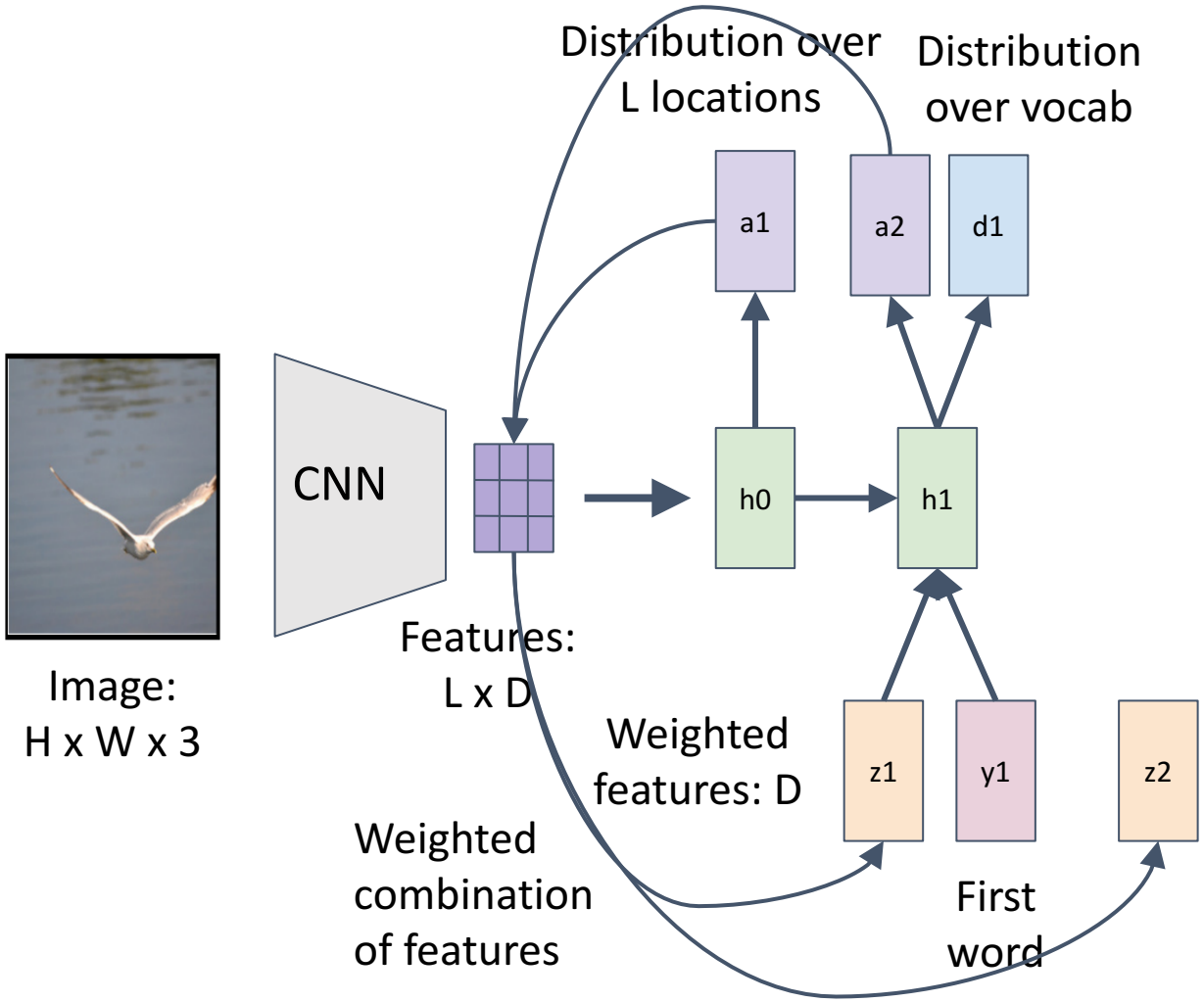
Soft Attention for Captioning



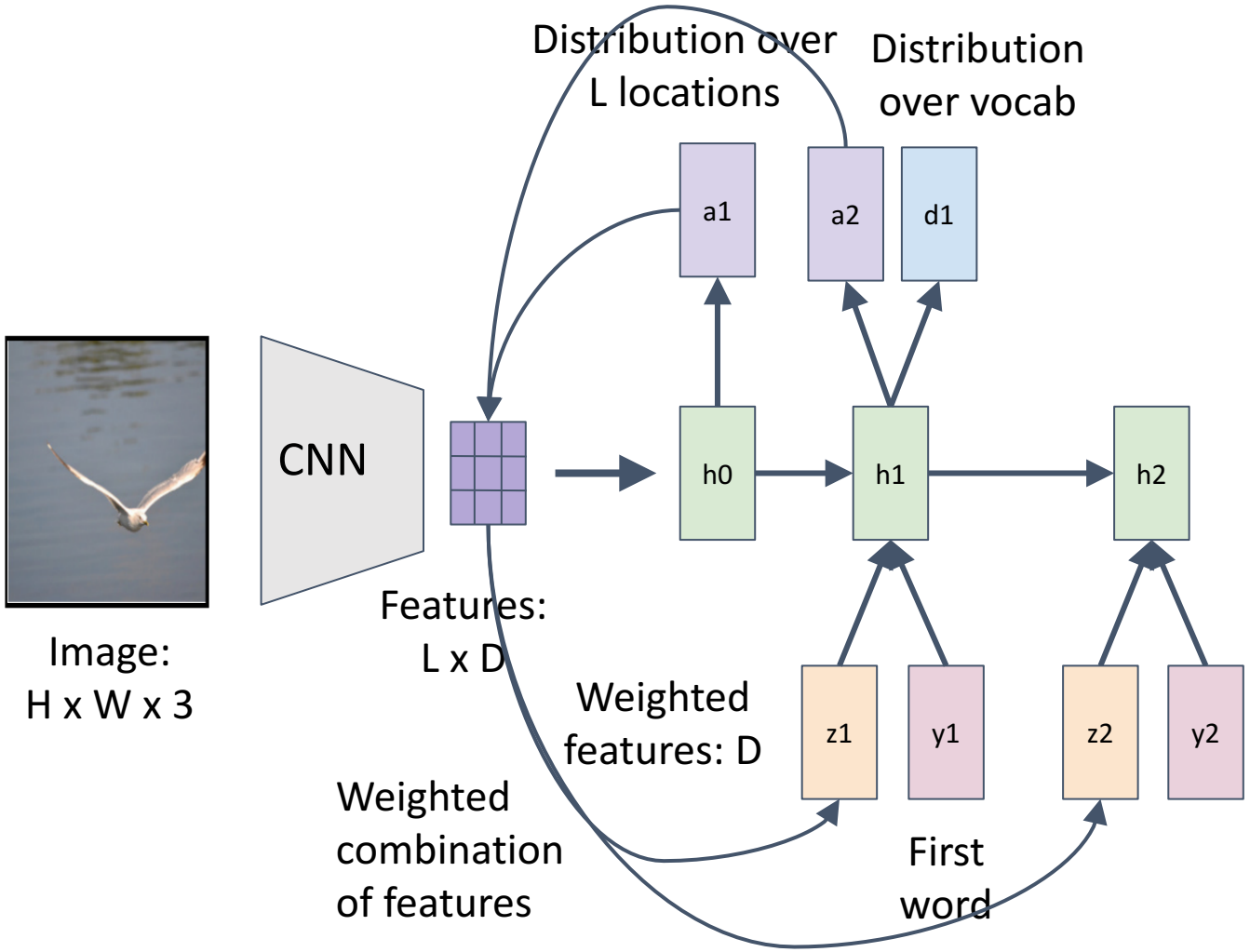
Soft Attention for Captioning



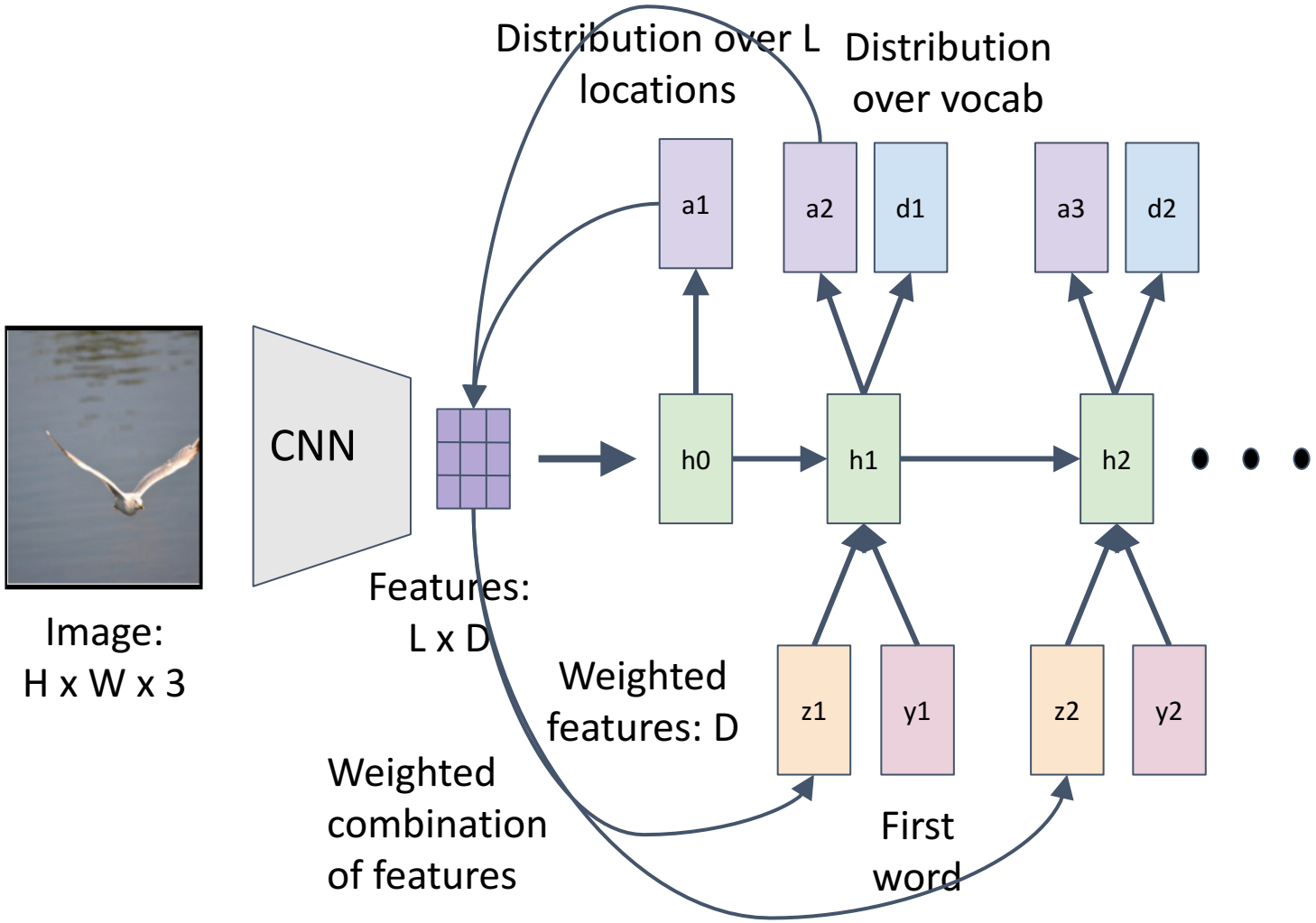
Soft Attention for Captioning



Soft Attention for Captioning

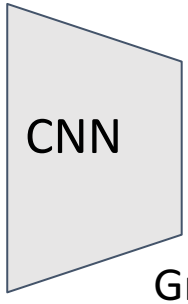


Soft Attention for Captioning



Soft vs Hard Attention

Image:
H x W x 3



a	b
c	d

Grid of features
(Each D-dimensional)

From
RNN:

p_a	p_b
p_c	p_d

Distribution over
grid locations
$p_a + p_b + p_c + p_c = 1$

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

Soft vs Hard Attention

Image:
H x W x 3

CNN

a	b
c	d

Grid of features
(Each D-dimensional)

Context vector z
(D-dimensional)

From
RNN:

p_a	p_b
p_c	p_d

Distribution over
grid locations
$p_a + p_b + p_c + p_c = 1$

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

Soft vs Hard Attention

Image:
H x W x 3

CNN

a	b
c	d

Grid of features
(Each D-dimensional)

Context vector z
(D-dimensional)

From
RNN:

p_a	p_b
p_c	p_d

Distribution over
grid locations
$p_a + p_b + p_c + p_c = 1$

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

Soft attention:
Summarize ALL locations
$z = p_a a + p_b b + p_c c + p_d d$

Derivative dz/dp is nice!
Train with gradient descent

Soft vs Hard Attention

Image:
H x W x 3

CNN

a	b
c	d

Grid of features
(Each D-dimensional)

Context vector z
(D-dimensional)

From
RNN:

p_a	p_b
p_c	p_d

Distribution over
grid locations
$p_a + p_b + p_c + p_c = 1$

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

Soft attention:
Summarize ALL locations
$z = p_a a + p_b b + p_c c + p_d d$

Derivative dz/dp is nice!
Train with gradient descent

Hard attention:
Sample ONE location
according to p, z = that vector

With argmax, dz/dp is zero
almost everywhere …
Can't use gradient descent;
need reinforcement learning

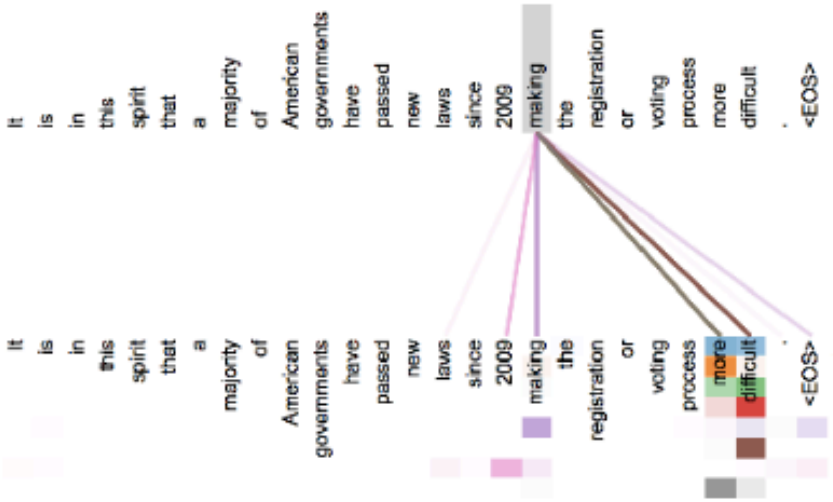
Multi-headed Attention

- **Idea:** multiple attention "heads" focus on different parts of the sentence

- e.g. Different heads for "copy" vs regular (Allamanis et al. 2016)

	Target	Attention Vectors		λ
m_1	set	$\alpha =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.012
		$\kappa =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	
m_2	use	$\alpha =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.974
		$\kappa =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	
m_3	browser	$\alpha =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.969
		$\kappa =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	
m_4	cache	$\alpha =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.583
		$\kappa =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	
m_5	END	$\alpha =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.066
		$\kappa =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	

- Or multiple independently learned heads (Vaswani et al. 2017)

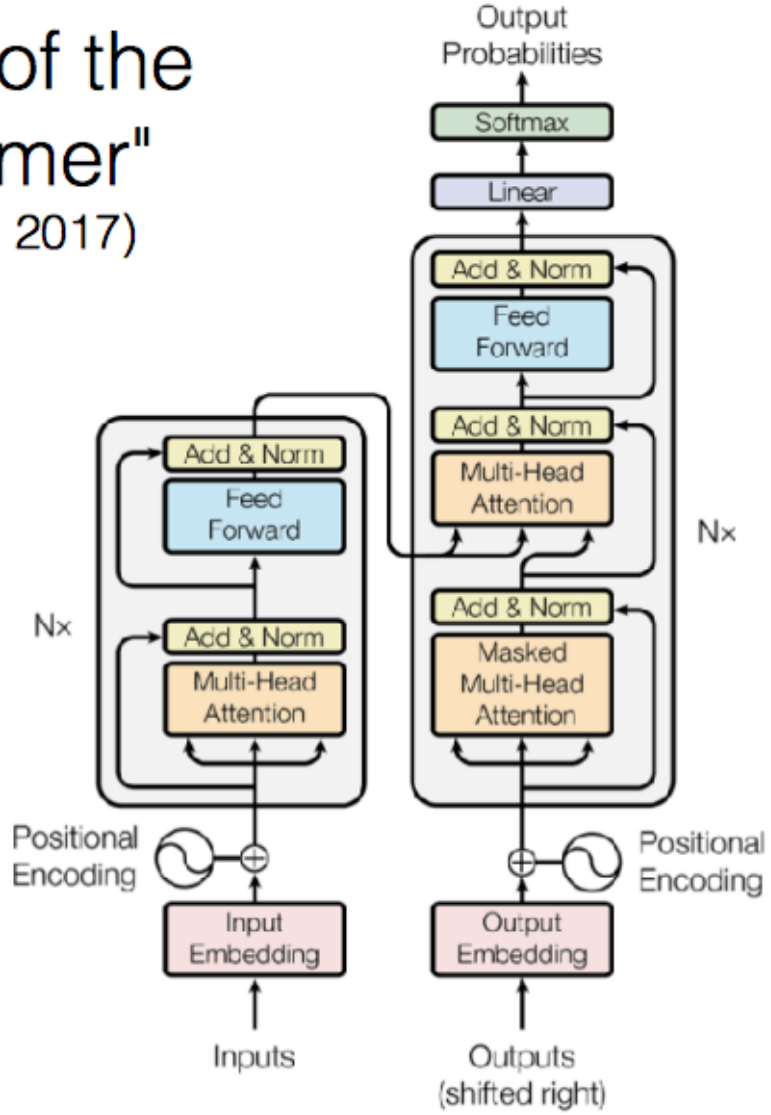


Attention is all you need

Summary of the "Transformer"

(Vaswani et al. 2017)

- A sequence-to-sequence model based entirely on attention
- Strong results on standard WMT datasets
- Fast: only matrix multiplications



Attention tricks

- **Self Attention:** Each layer combines words with others
- **Multi-headed Attention:** 8 attention heads learned independently
- **Normalized Dot-product Attention:** Remove bias in dot product when using large networks
- **Positional Encodings:** Make sure that even if we don't have RNN, can still distinguish positions

Attention Takeaways

Performance:

- Attention models can ***improve accuracy*** and ***reduce computation*** at the same time.

Complexity:

- There are many design choices.
- Those choices have a big effect on performance.
- Ensembling has unusually large benefits.
- Simplify where possible!

Attention Takeaways

Explainability:

- Attention models encode explanations.
- Both locus and trajectory help understand what's going on.

Hard vs. Soft:

- Soft models are easier to train, hard models require reinforcement learning.
- They can be combined, as in Luong et al.