DSC250: Advanced Data Mining

Language Models

Zhiting Hu
Lecture 9, October 26, 2023

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Last lecture

- Neural language models:
  - Embedding: one-hot vectors -> embedding vectors
  - Neural networks

Neural Architectures of LMs

Outline

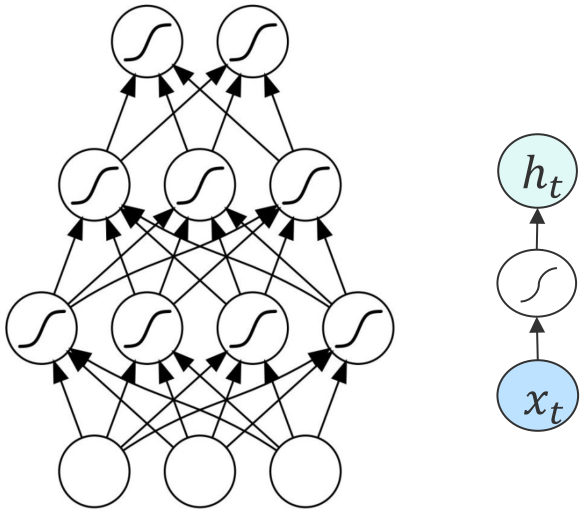
- Recurrent Networks (RNNs)
  - Long-range dependency, vanishing gradients
  - LSTM
  - RNNs in different forms
- Attention Mechanisms
  - (Query, Key, Value)
  - Attention on Text and Images
- Transformers: Multi-head Attention
  - Transformer
  - BERT

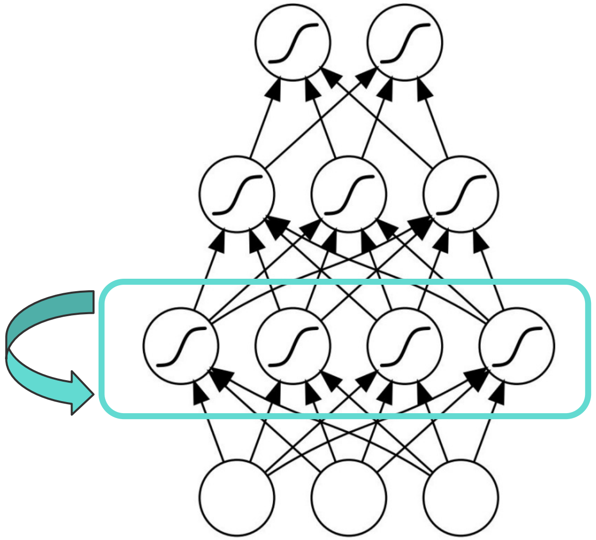
Outline

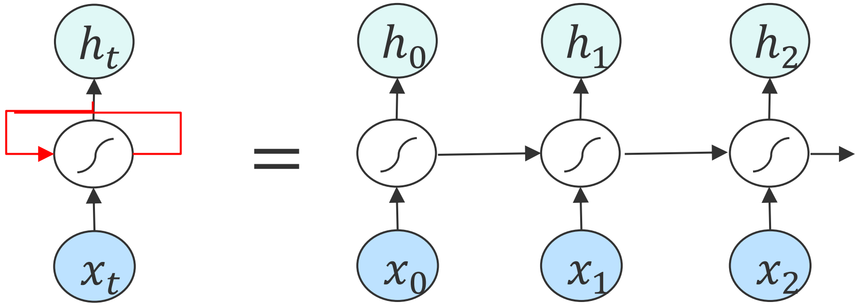
- Recurrent Networks (RNNs)
  - Long-range dependency, vanishing gradients
  - LSTM
  - RNNs in different forms
- Attention Mechanisms
  - (Query, Key, Value)
  - Attention on Text and Images
- Transformers: Multi-head Attention
  - Transformer
  - BERT

ConvNets v.s. Recurrent Networks (RNNs)

- Spatial Modeling *vs.* Sequential Modeling
- Fixed *vs.* variable number of computation steps.







The output depends ONLY on the current input

The hidden layers and the output additionally depend on previous states of the hidden layers

RNNs in Various Forms

One to One

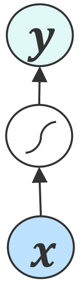


Image
classification

One to Many

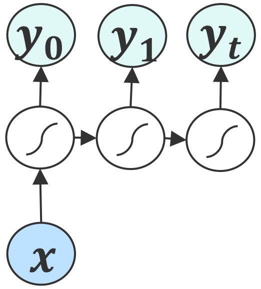
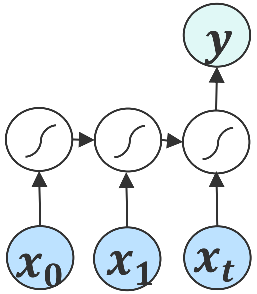


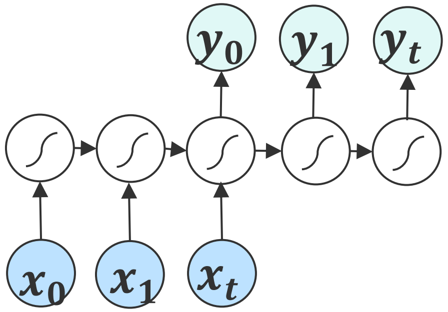
Image
captioning

Many to One



Sentence sentiment analysis /
Video recognition

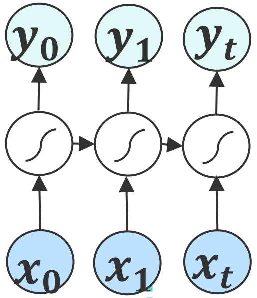
Many to Many



Machine Translation

(Sequence-to-sequence)

Many to Many

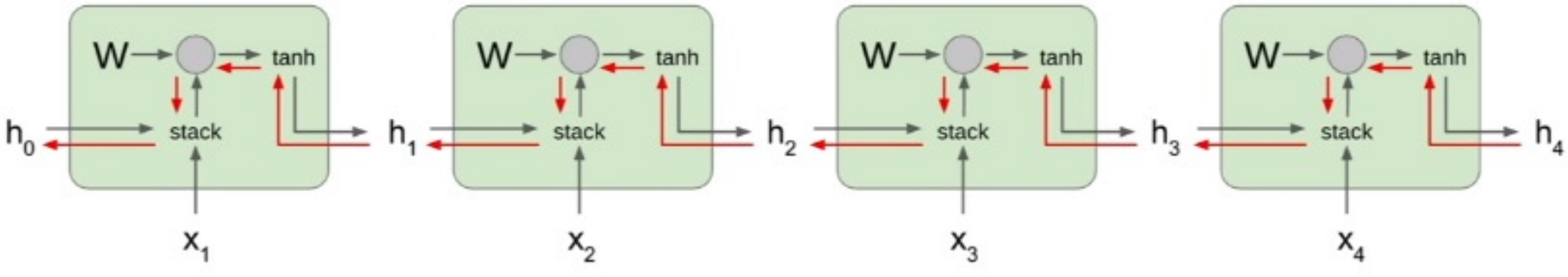


Named Entity Recognition

(Sequence tagging)

Vanishing / Exploding Gradients in RNNs

$$\boldsymbol{h}_t = tanh(W^{hh}\boldsymbol{h}_{t-1} + W^{hx}\boldsymbol{x}_t)$$



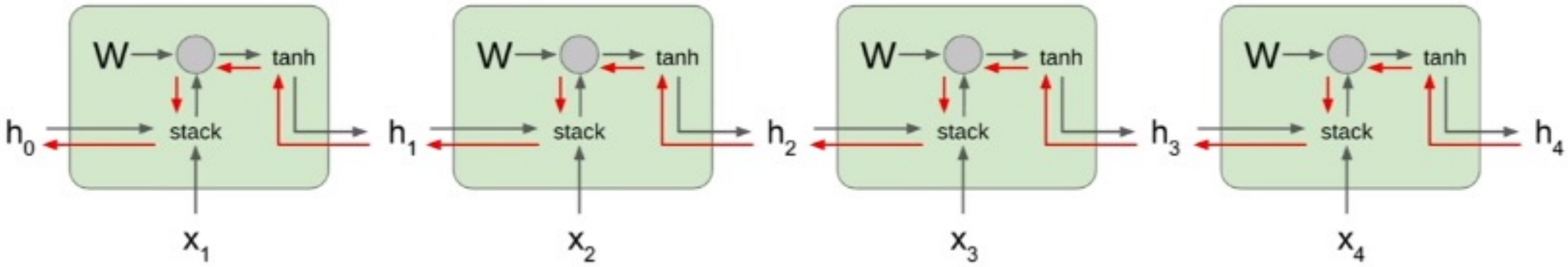
Bengio et al., 1994 “Learning long-term dependencies with gradient descent is difficult”

Source: CS231N Stanford

Pascanu et al., 2013 “On the difficulty of training recurrent neural networks”

Vanishing / Exploding Gradients in RNNs

$$\boldsymbol{h}_t = tanh(W^{hh}\boldsymbol{h}_{t-1} + W^{hx}\boldsymbol{x}_t)$$



Computing gradient of h_0 involves many factors of W (and repeated tanh)

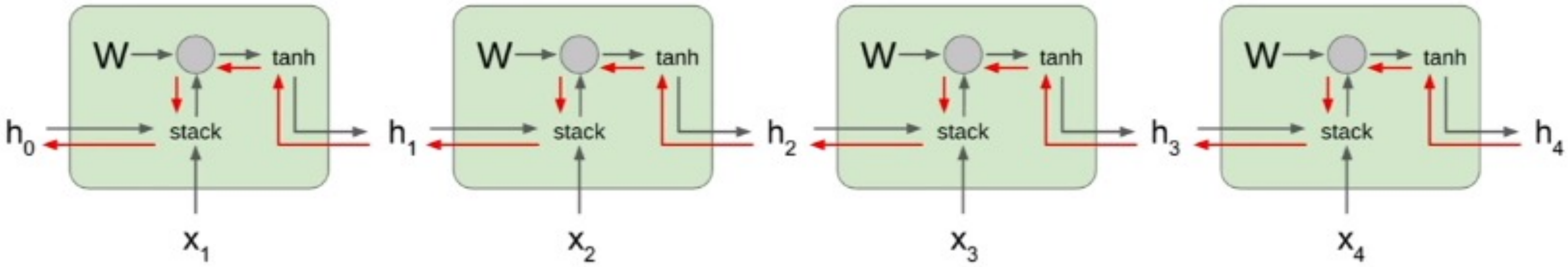
Bengio et al., 1994 "Learning long-term dependencies with gradient descent is difficult"

Pascanu et al., 2013 "On the difficulty of training recurrent neural networks"

Source: CS231N Stanford

Vanishing / Exploding Gradients in RNNs

$$\boldsymbol{h}_t = tanh(W^{hh}\boldsymbol{h}_{t-1} + W^{hx}\boldsymbol{x}_t)$$



Computing gradient of h_0 involves many factors of W (and repeated tanh)

Largest singular value > 1:
Exploding gradients

→ **Gradient clipping**: Scale gradient if its norm is too big

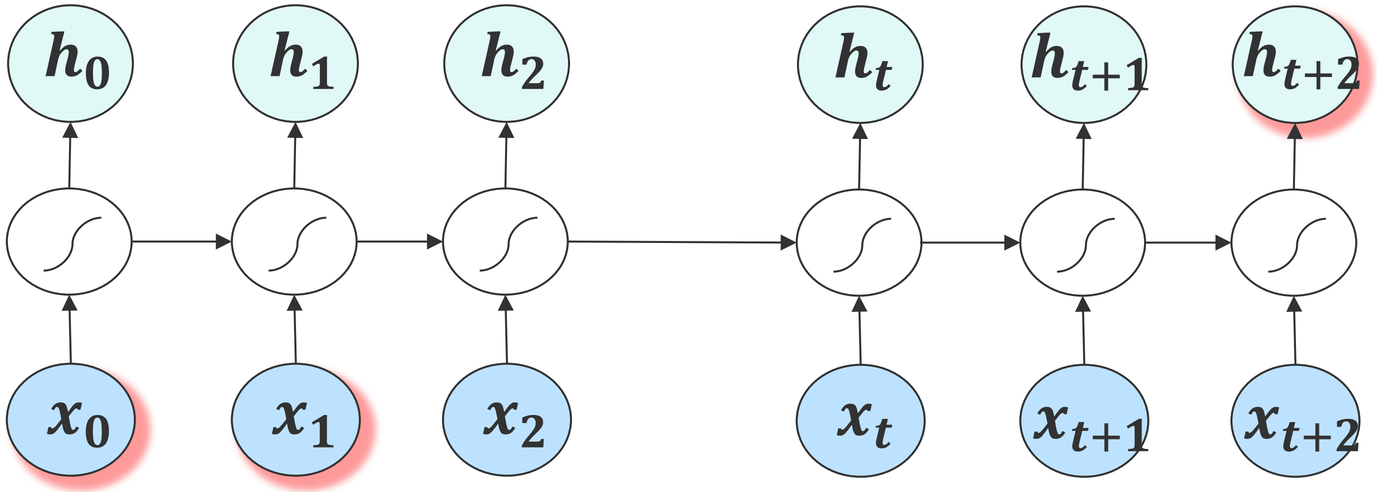
```
grad_norm = np.sum(grad * grad)
if grad_norm > threshold:
    grad *= (threshold / grad_norm)
```

Largest singular value < 1:
Vanishing gradients

Source: CS231N Stanford

Bengio et al., 1994 “Learning long-term dependencies with gradient descent is difficult”
Pascanu et al., 2013 “On the difficulty of training recurrent neural networks”

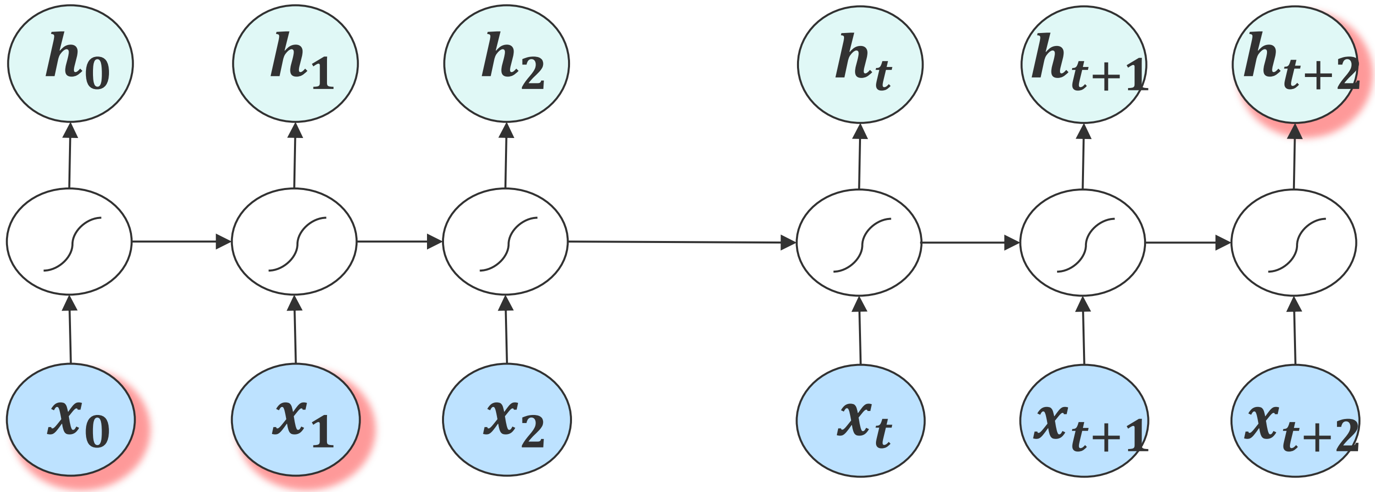
Long-term Dependency Problem



I live in France and I know ___________

Example courtesy: Manik Soni

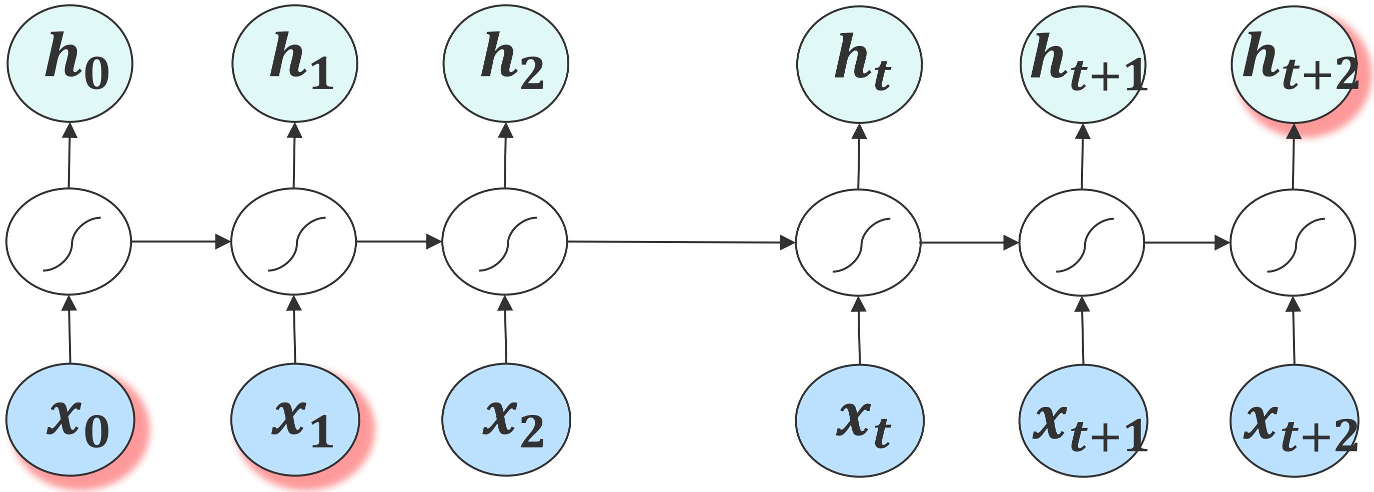
Long-term Dependency Problem



I live in France and I know ___*French*___

Example courtesy: Manik Soni

Long-term Dependency Problem



I live in France and I know __*French*__

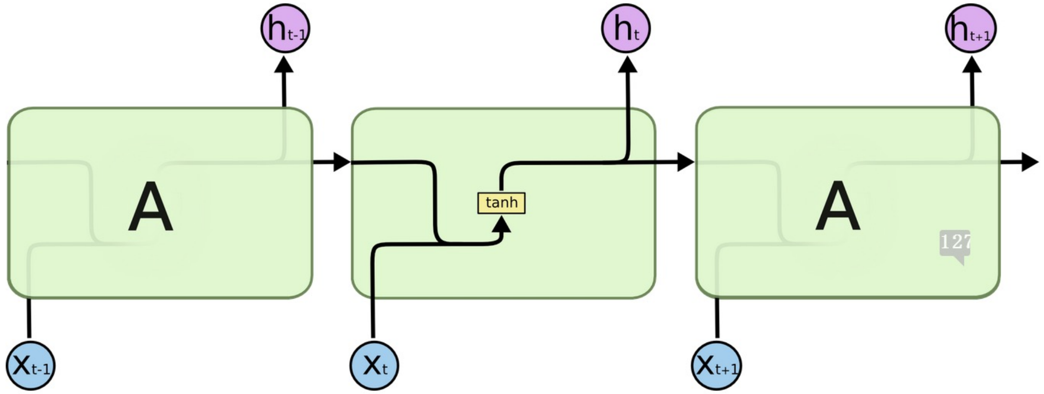
I live in France, a beautiful country, and I know __*French*__

Example courtesy: Manik Soni

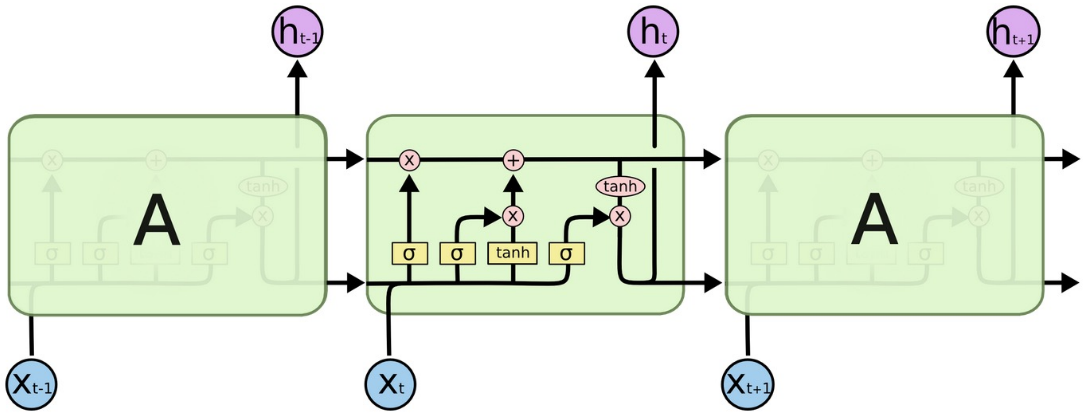
Long Short Term Memory (LSTM)

- LSTMs are designed to explicitly alleviate the long-term dependency problem [Horchreiter & Schmidhuber (1997)]

Standard RNN

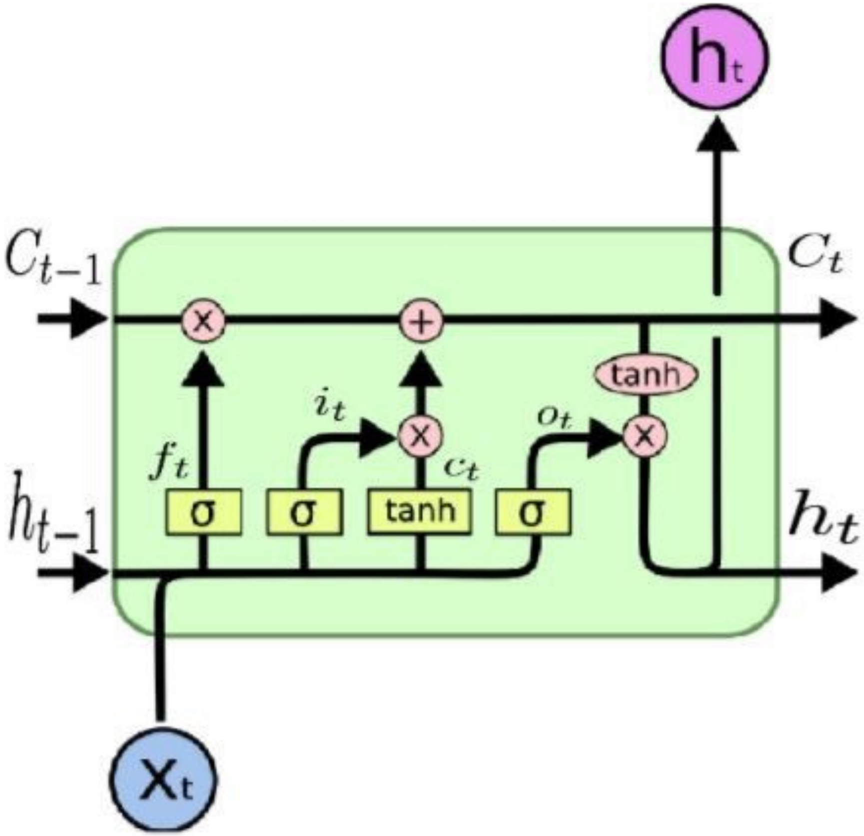


LSTM



Long Short Term Memory (LSTM)

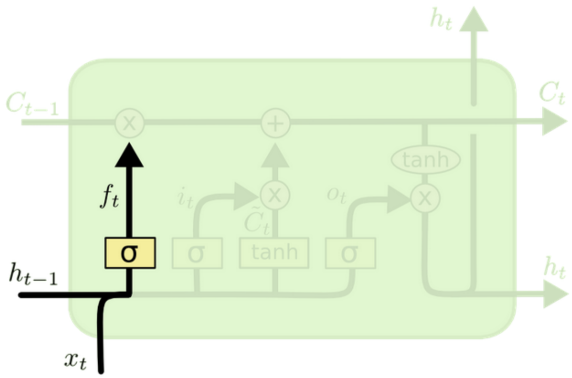
- Gate functions make decisions of reading, writing, and resetting information



- Forget gate: whether to erase cell (reset)
- Input gate: whether to write to cell (write)
- Output gate: how much to reveal cell (read)

Long Short Term Memory (LSTM)

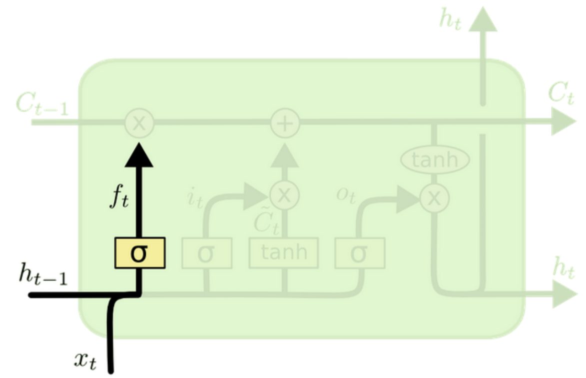
- **Forget gate:** decides what must be removed from $\boldsymbol{h}_{t-1}$



$$\boldsymbol{f_t} = \boldsymbol{\sigma}(\boldsymbol{W_f} \cdot [\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_f})$$

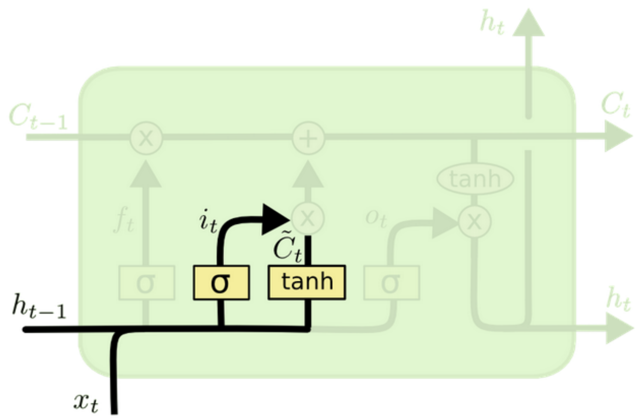
Long Short Term Memory (LSTM)

- **Forget gate:** decides what must be removed from $\boldsymbol{h_{t-1}}$



$$\boldsymbol{f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)}$$

- **Input gate:** decides what new information to store in the cell

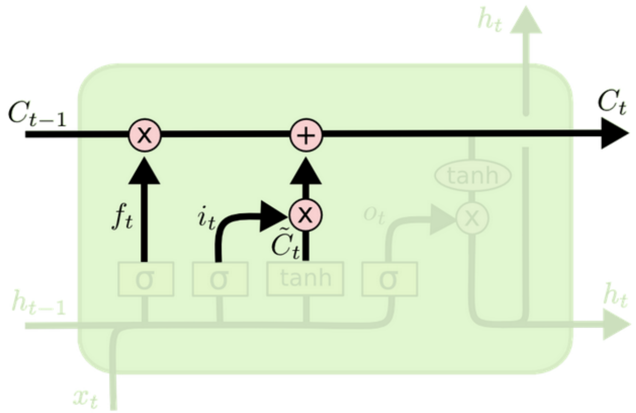


$$\boldsymbol{i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)}$$

$$\boldsymbol{\tilde{C}_t} = \tanh(\boldsymbol{W_C \cdot [h_{t-1}, x_t] + b_C})$$

Long Short Term Memory (LSTM)

- Update cell state:



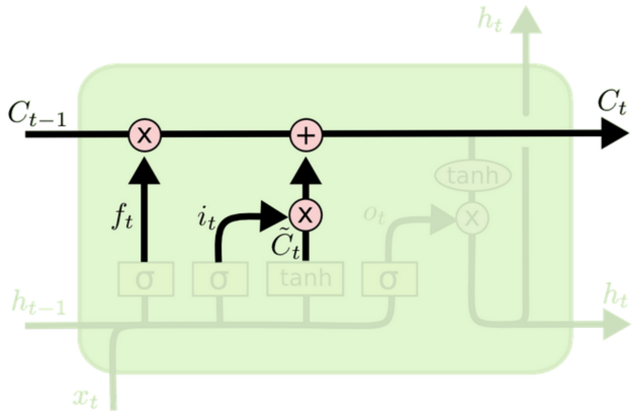
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

forgetting unneeded things

scaling the new candidate values by how much we decided to update each state value.

Long Short Term Memory (LSTM)

- Update cell state:

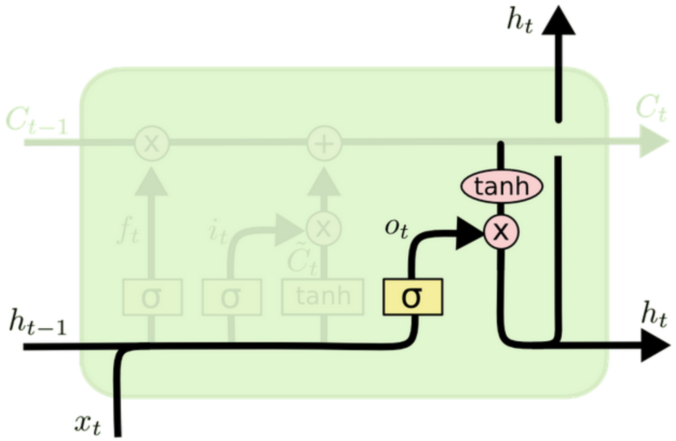


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

forgetting unneeded things

scaling the new candidate values by how much we decided to update each state value.

- **Output gate:** decides what to output from our cell state

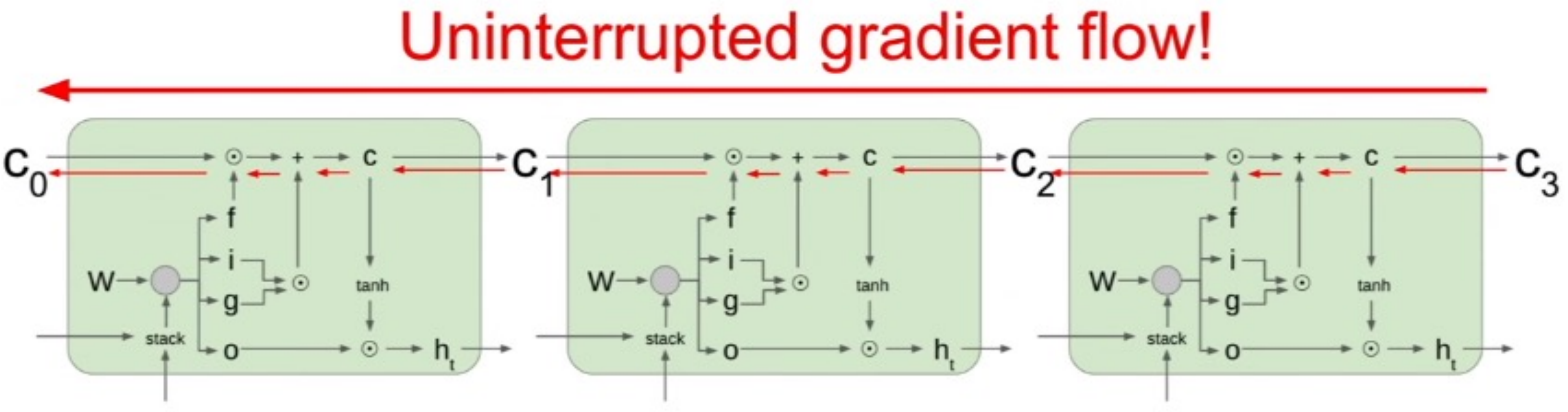


$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

sigmoid decides what parts of the cell state we're going to output

Backpropagation in LSTM



- No multiplication with matrix W during backprop
- Multiplied by different values of forget gate -> less prone to vanishing/exploding gradient

Source: CS231N Stanford

RNNs in Various Forms

One to One

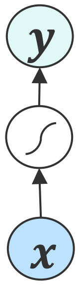


Image classification

One to Many

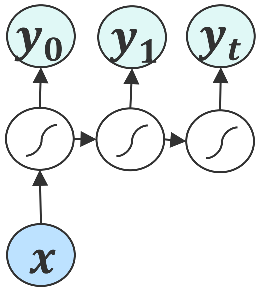
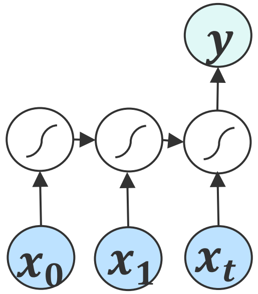


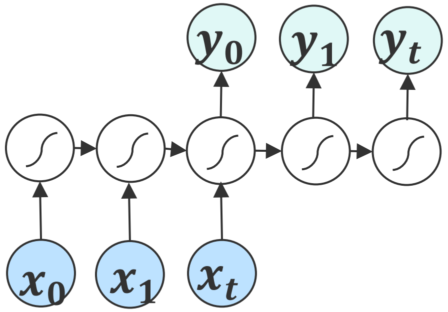
Image captioning

Many to One



Sentence sentiment analysis / Video recognition

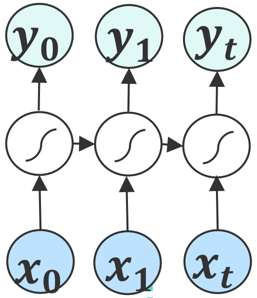
Many to Many



Machine Translation

(Sequence-to-sequence)

Many to Many

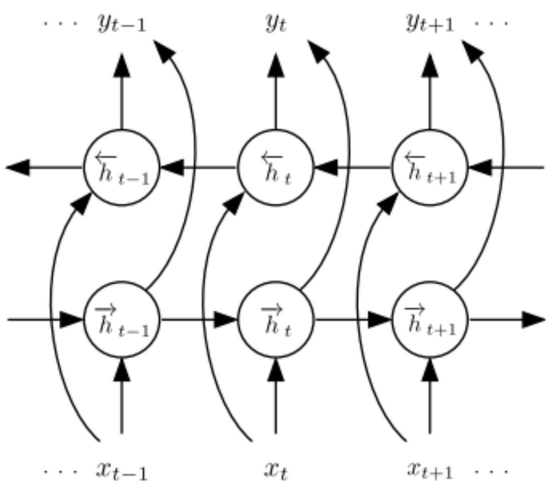


Named Entity Recognition

(Sequence tagging)

RNNs in Various Forms

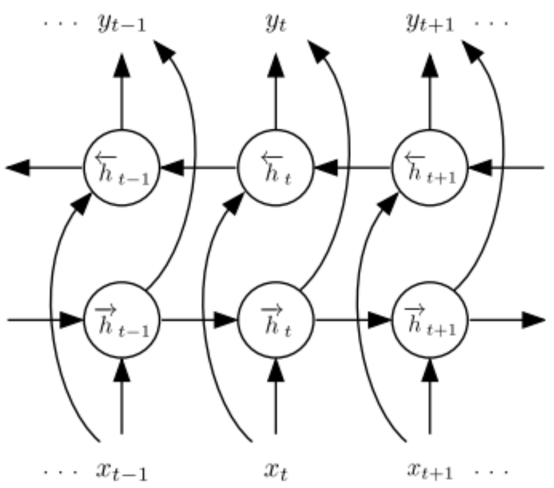
- Bi-directional RNN
  - Hidden state is the concatenation of both forward and backward hidden states.
  - Allows the hidden state to capture both past and future information.



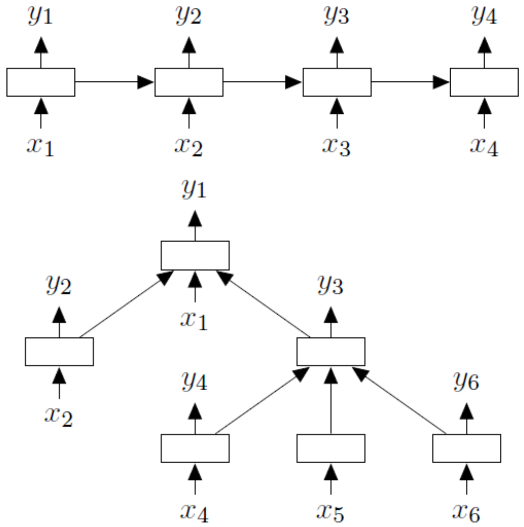
[Speech Recognition with Deep Recurrent Neural Networks, Alex Graves]

RNNs in Various Forms

- Bi-directional RNN
  - Hidden state is the concatenation of both forward and backward hidden states.
  - Allows the hidden state to capture both past and future information.
- Tree-structured RNN
  - Hidden states condition on both an input vector and the hidden states of arbitrarily many child units.
  - Standard LSTM = a special case of tree-LSTM where each internal node has exactly one child.



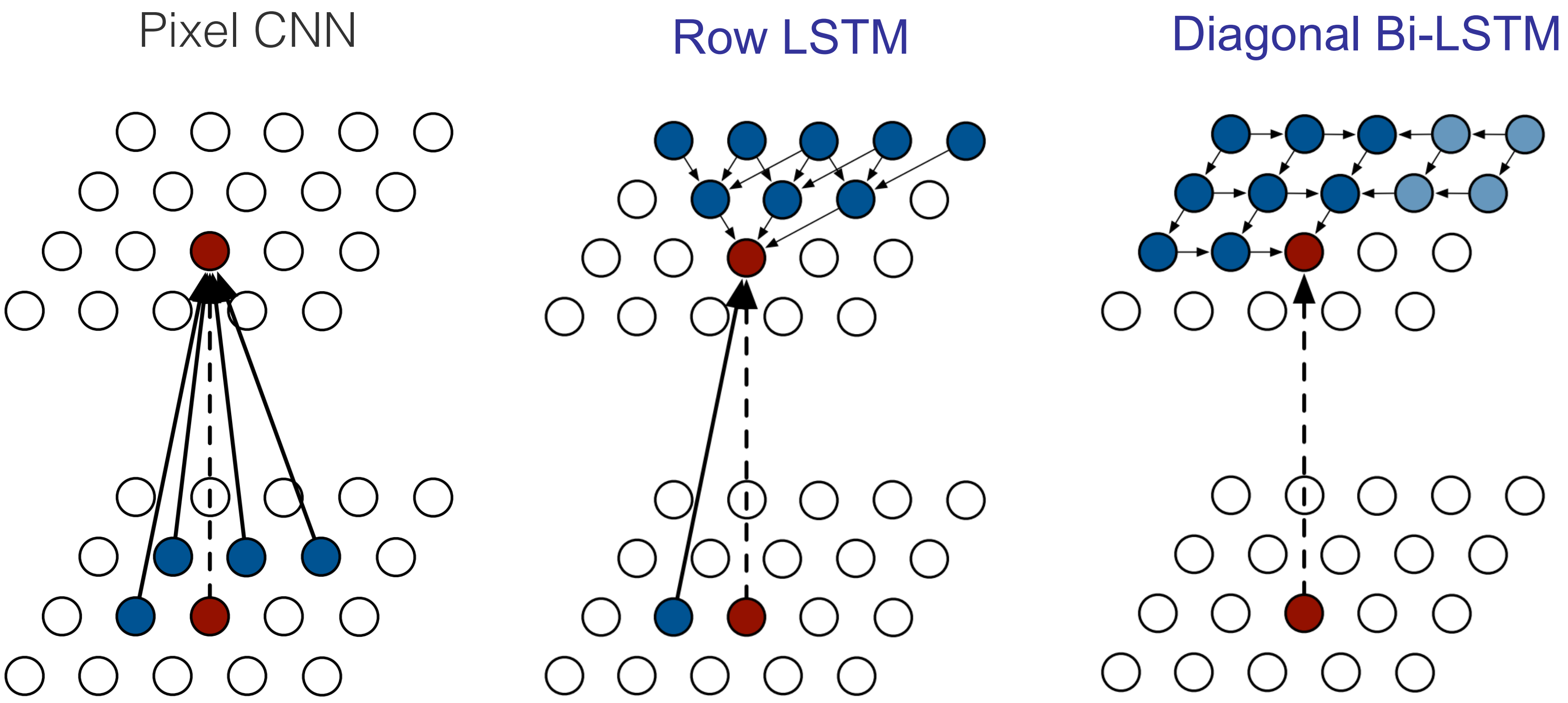
[Speech Recognition with Deep Recurrent Neural Networks, Alex Graves]



Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, Tai. et al.

RNNs in Various Forms

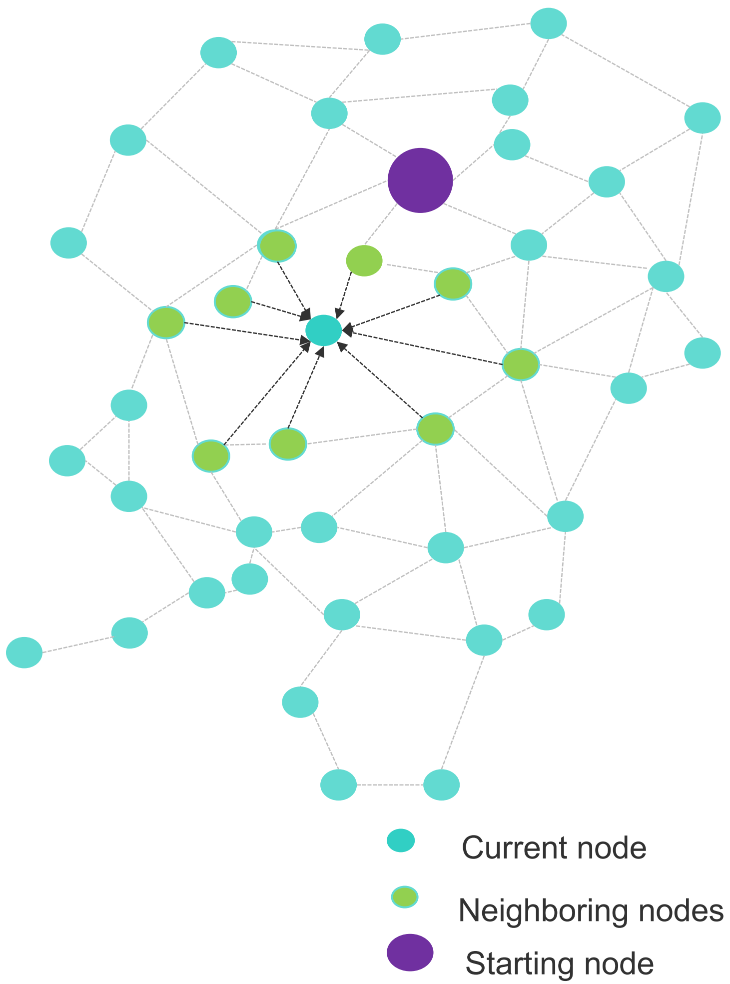
- RNN for 2-D sequences



[Pixel Recurrent Neural Networks, van den Oord. et al. 2016]

RNNs in Various Forms

- RNN for Graph Structures
  - Used in, e.g., image segmentation



[Semantic Object Parsing with Graph LSTM. Liang et al. 2016]

Outline

- Recurrent Networks (RNNs)
  - Long-range dependency, vanishing gradients
  - LSTM
  - RNNs in different forms
- Attention Mechanisms
  - (Query, Key, Value)
  - Attention on Text and Images
- Transformers: Multi-head Attention
  - Transformer
  - BERT

Attention: Examples

- Chooses which features to pay attention to

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.

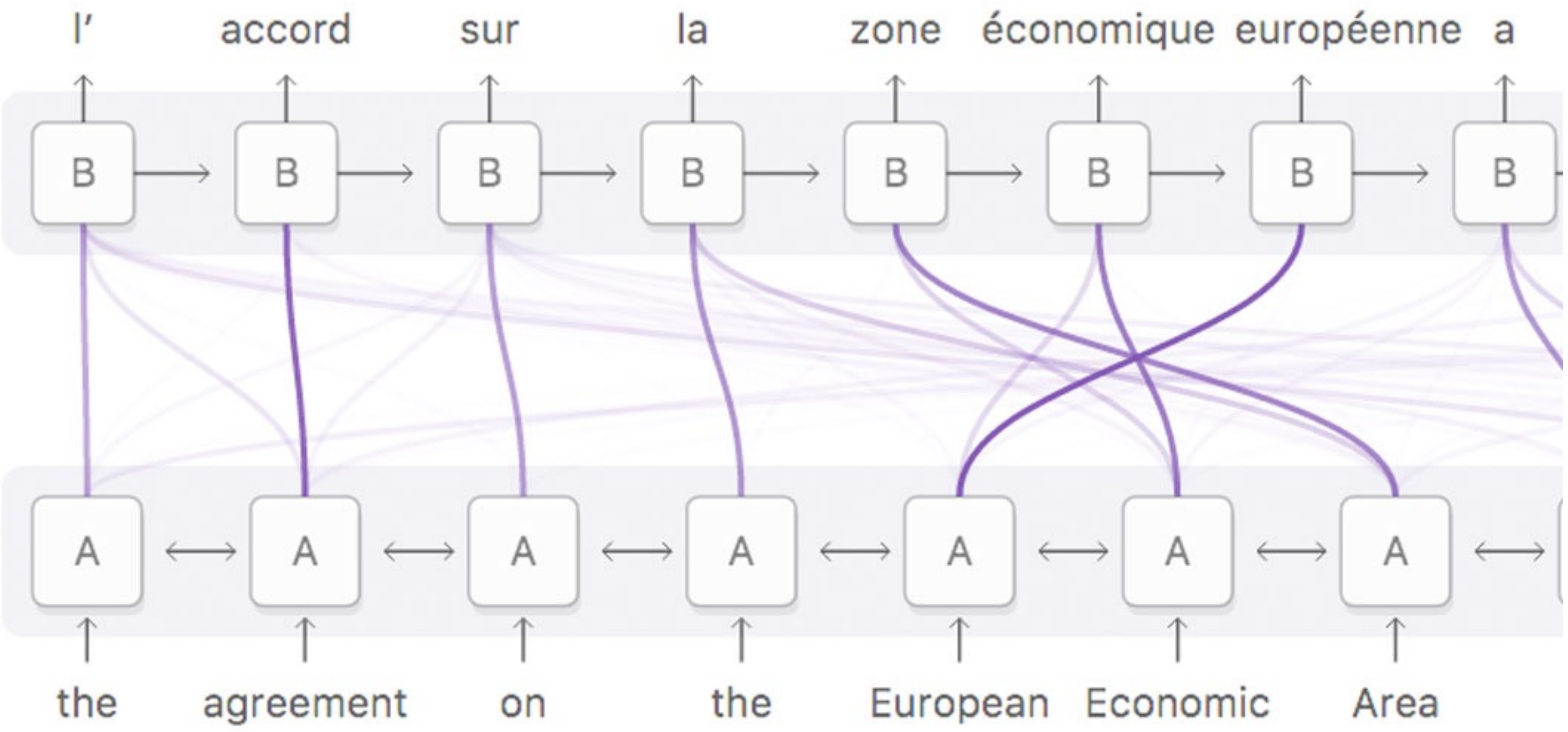
A group of people sitting on a boat in the water.

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

Image captioning [Show, attend and tell. Xu et al. 15]

Attention: Examples

- Chooses which features to pay attention to



Machine Translation

Figure courtesy: Olah & Carter, 2016

Why Attention?

Figure courtesy: keitakurita

Why Attention?

- Long-range dependencies
  - Dealing with gradient vanishing problem

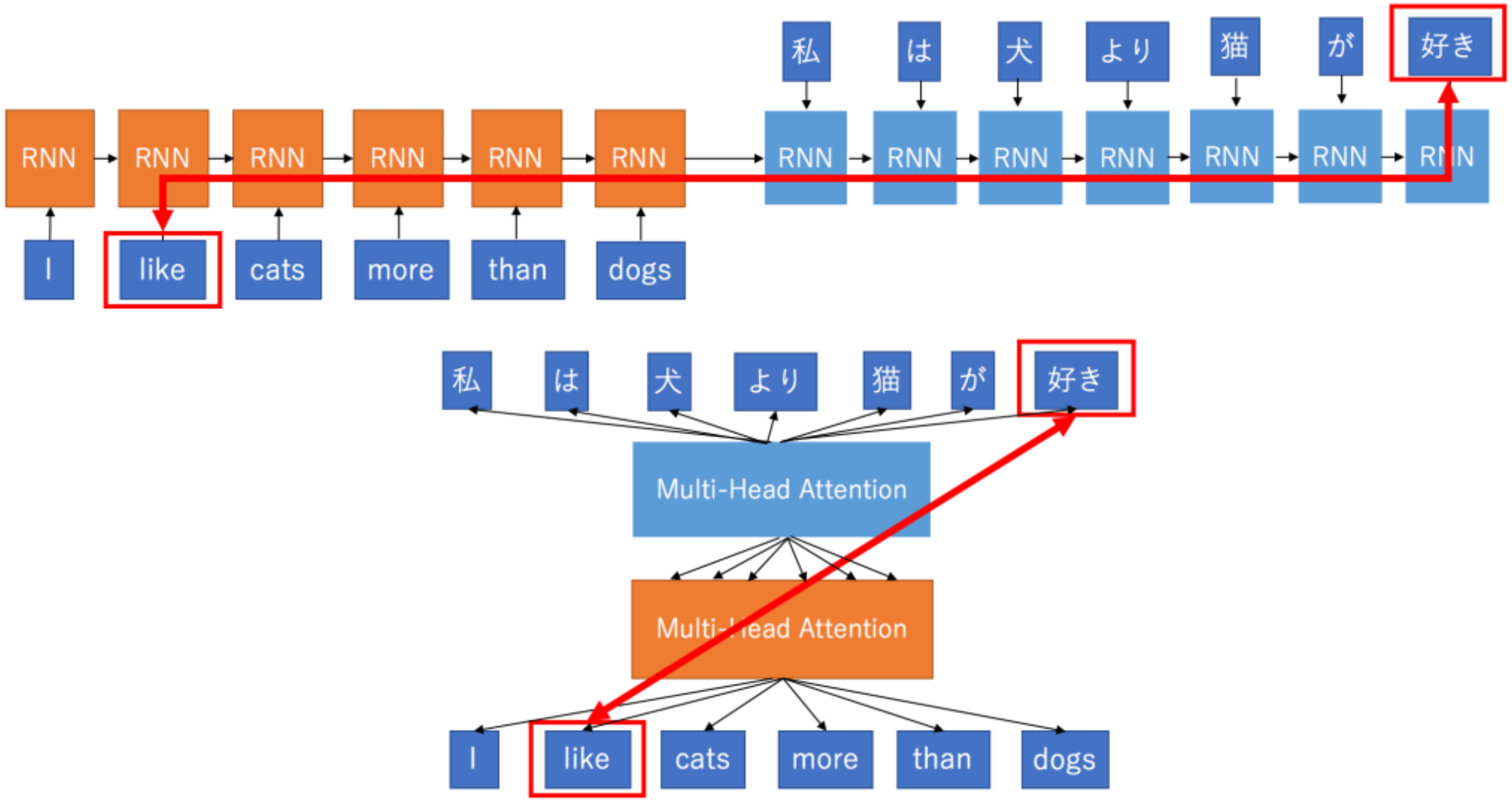


Figure courtesy: keitakurita

Why Attention?

- Long-range dependencies
  - Dealing with gradient vanishing problem
- Fine-grained representation instead of a single global representation
  - Attending to smaller parts of data: patches in images, words in sentences

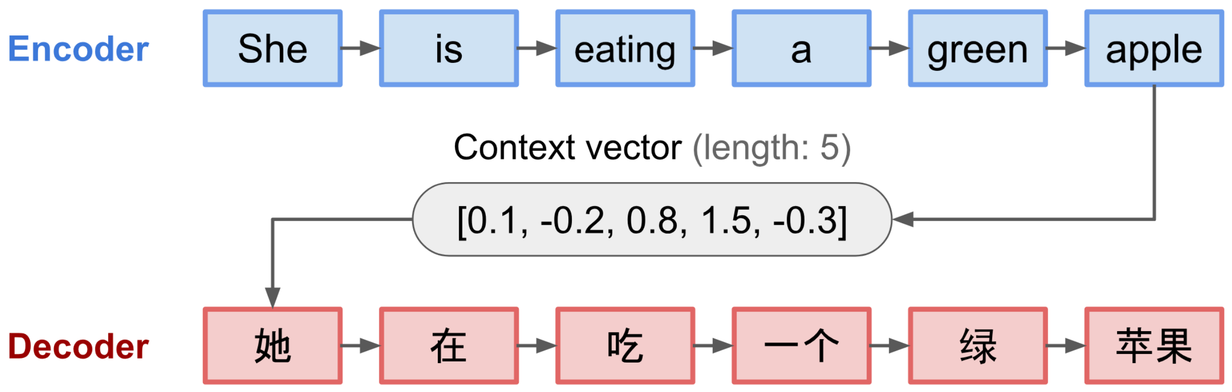


Figure courtesy: Lilian Weng

Why Attention?

- Long-range dependencies
  - Dealing with gradient vanishing problem
- Fine-grained representation instead of a single global representation
  - Attending to smaller parts of data: patches in images, words in sentences
- Improved Interpretability

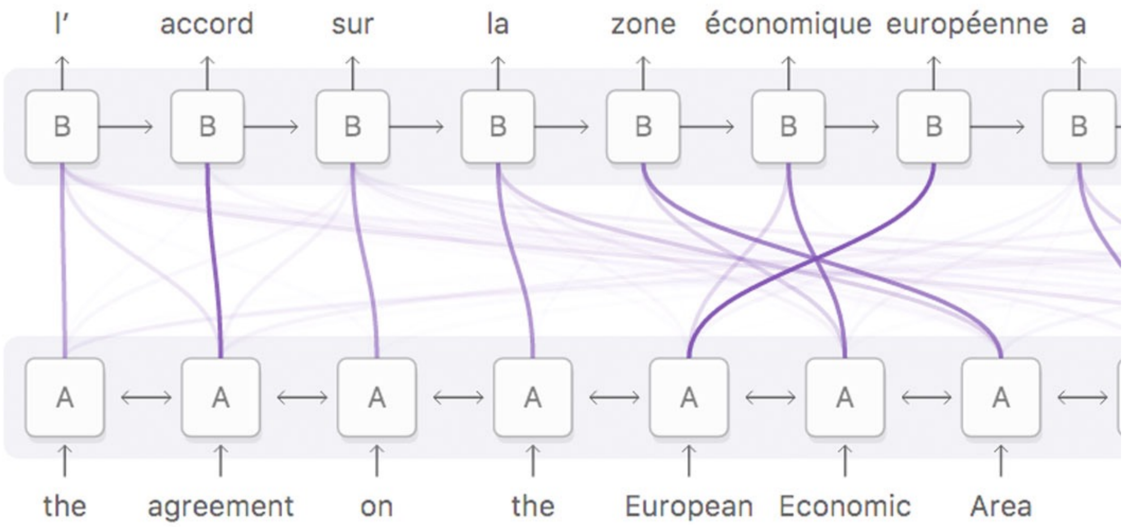


Figure courtesy: Olah & Carter, 2016

Attention Computation

- Encode each token in the input sentence into vectors
- When decoding, perform a linear combination of these vectors, weighted by "attention weights"
  - $\boldsymbol{a} = \text{softmax}(\boldsymbol{alignment\_scores})$

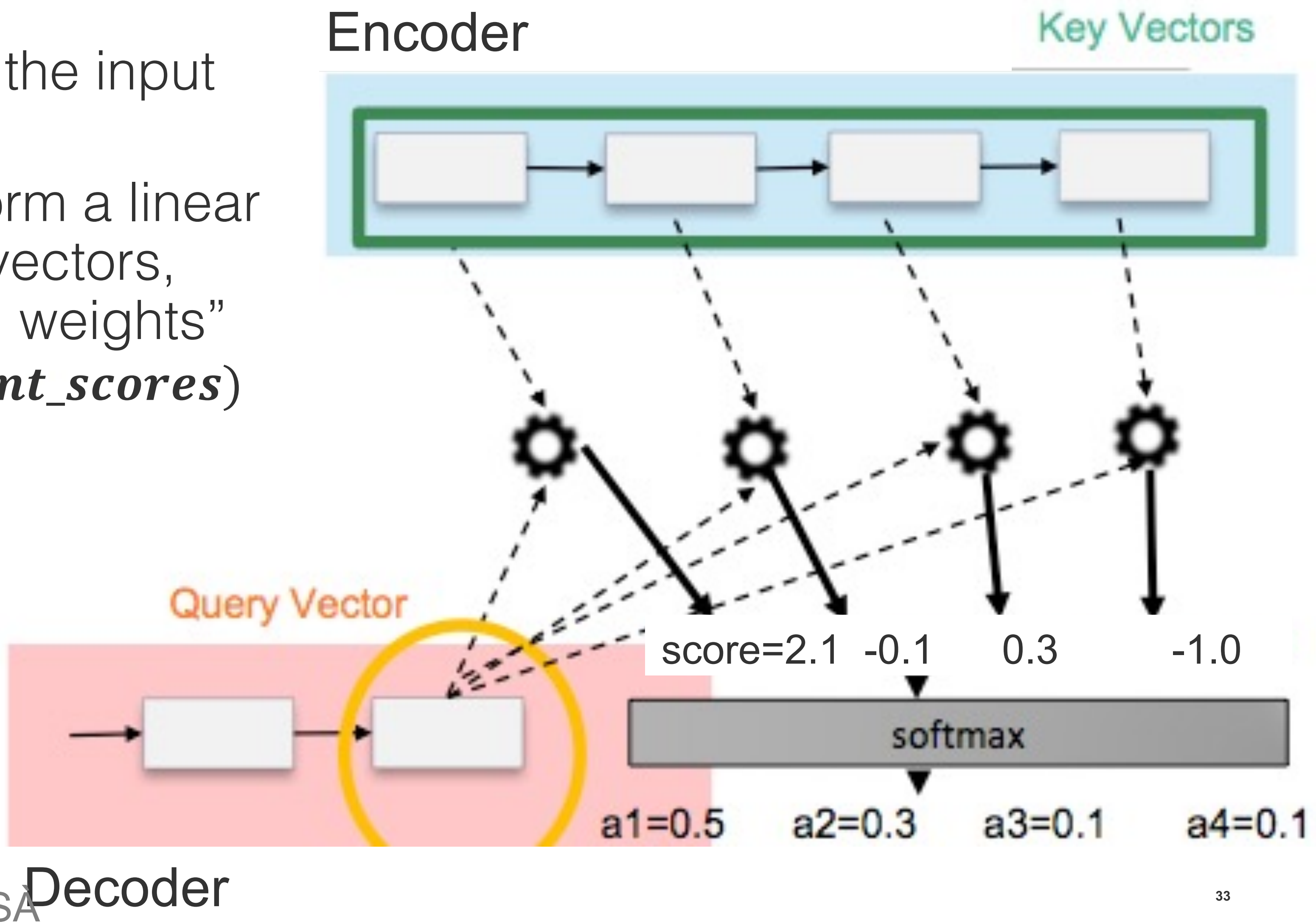
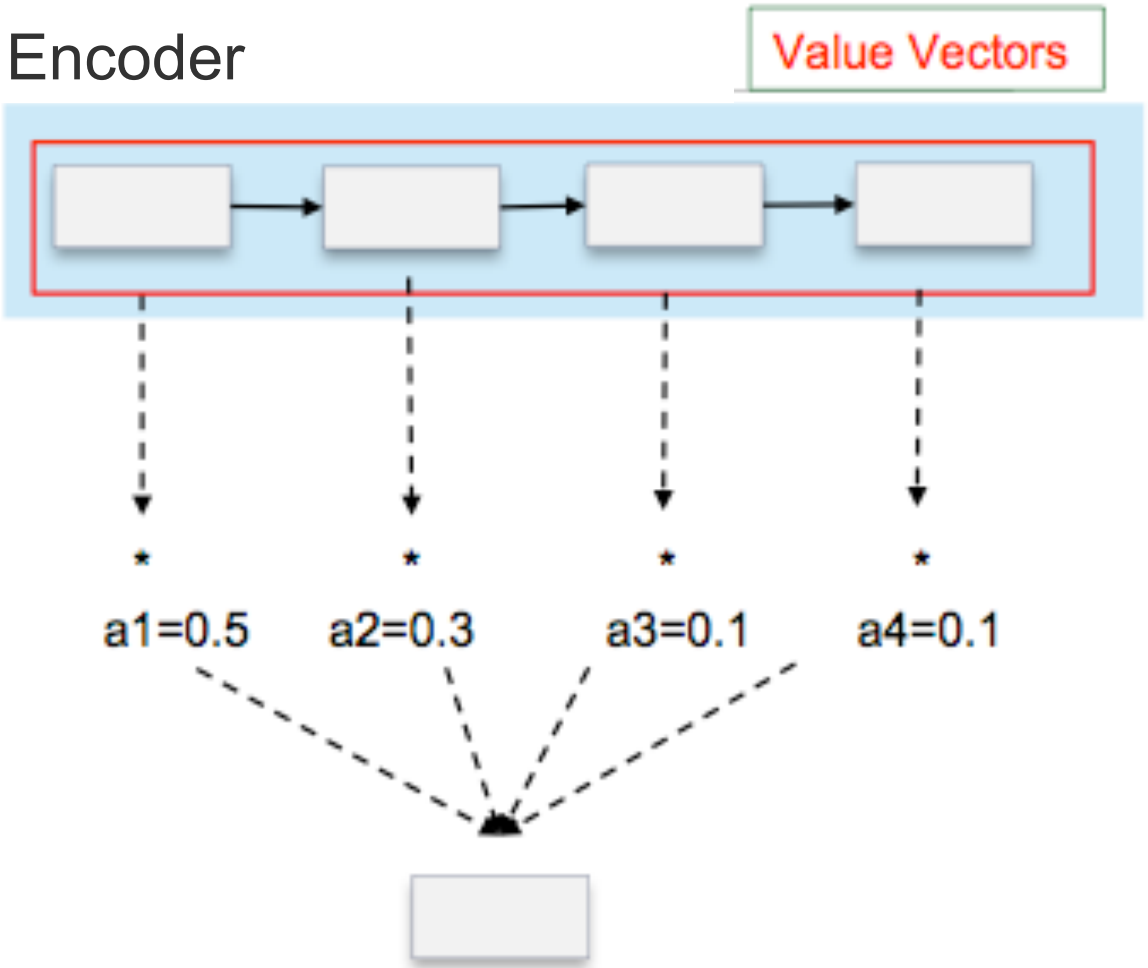


Figure courtesy: MARTA R. COSTA-JUSSÀ

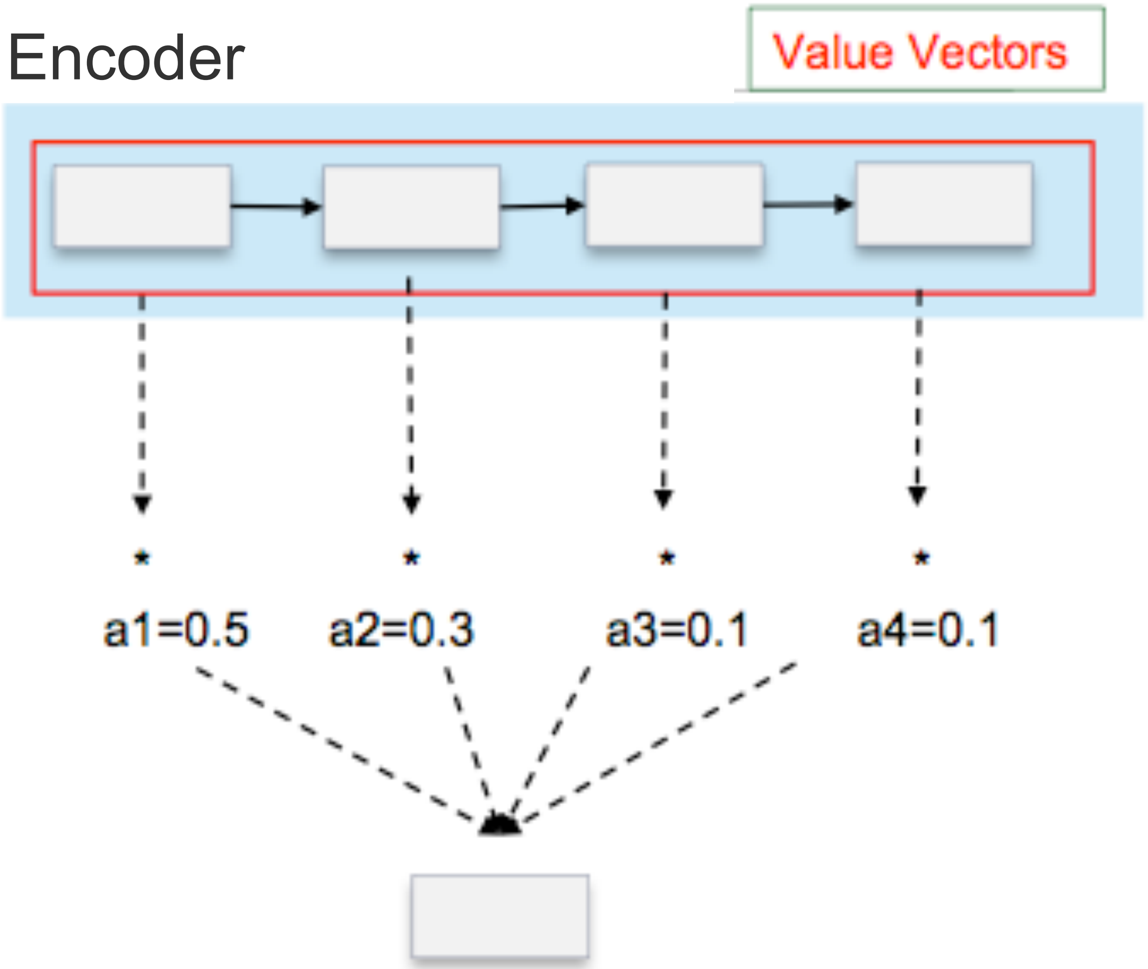
Attention Computation (cont'd)

- Combine together value by taking the weighted sum



Attention Computation (cont'd)

- Combine together value by taking the weighted sum



- **Query**: decoder state
- **Key**: all encoder states
- **Value**: all encoder states

Attention Variants

- Popular attention mechanisms with different alignment score functions

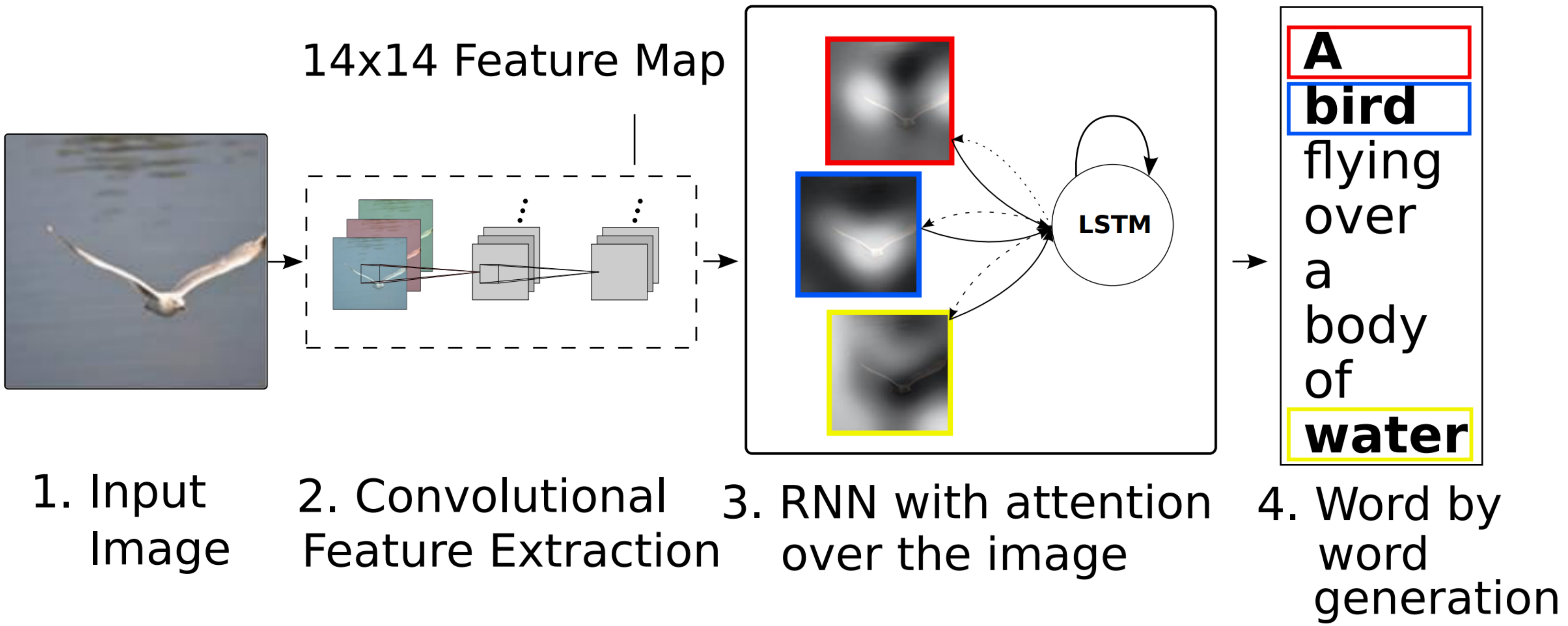
Alignment score = f(Query, Keys)

- Query: decoder state s_t
- Key: all encoder states h_i
- Value: all encoder states h_i

Name	Alignment score function	Citation
Content-base attention	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$	Graves2014
Additive(*)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$	Bahdanau2015
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \boldsymbol{s}_t)$ Note: This simplifies the softmax alignment to only depend on the target position.	Luong2015
General	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix in the attention layer.	Luong2015
Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{s}_t^\top \boldsymbol{h}_i$	Luong2015
Scaled Dot-Product(^)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\mathbf{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.	Vaswani2017

Courtesy: Lilian Weng

Attention on Images – Image Captioning

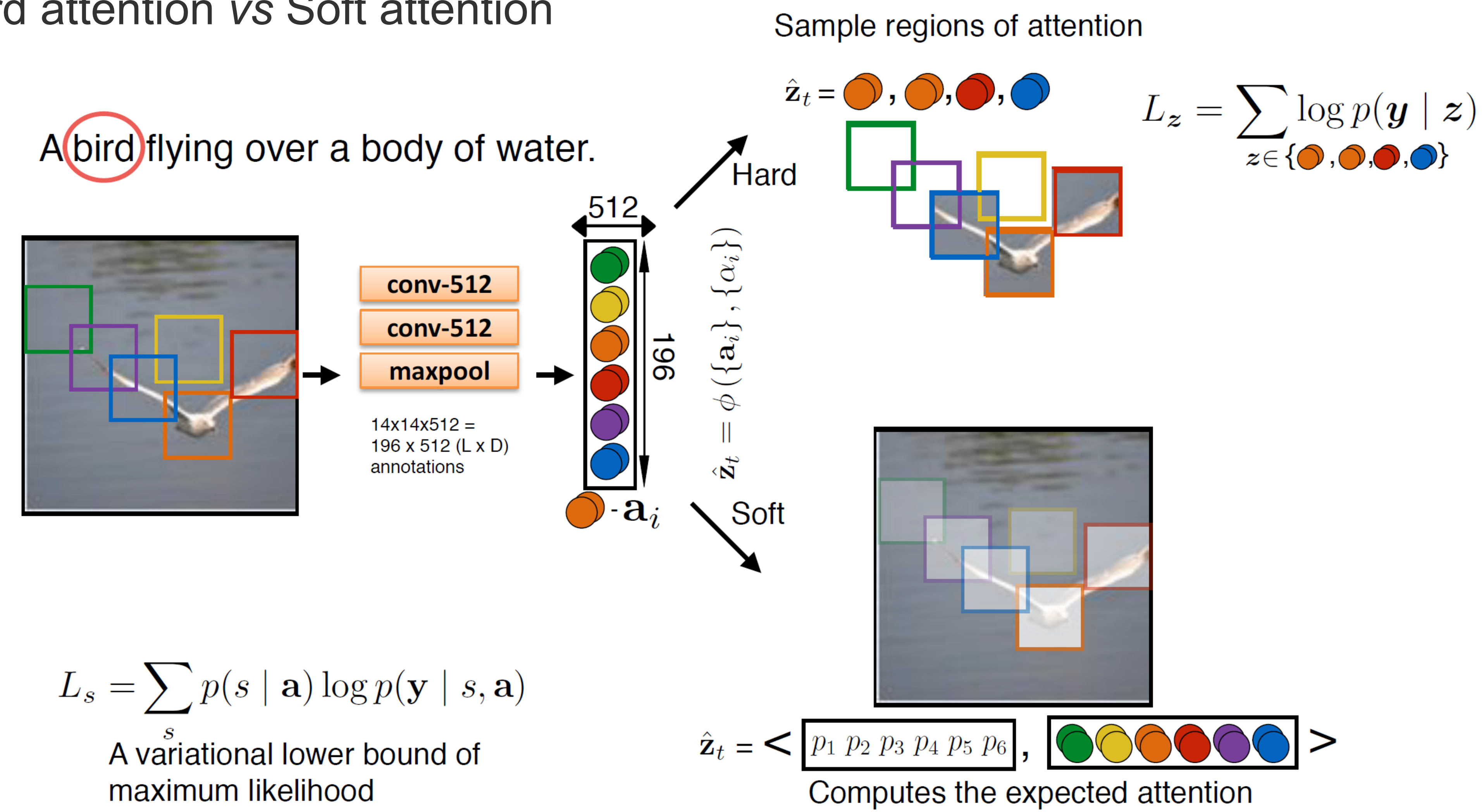


- Query: decoder state
- Key: visual feature maps
- Value: visual feature maps

[Show, attend and tell. Xu et al. 15]

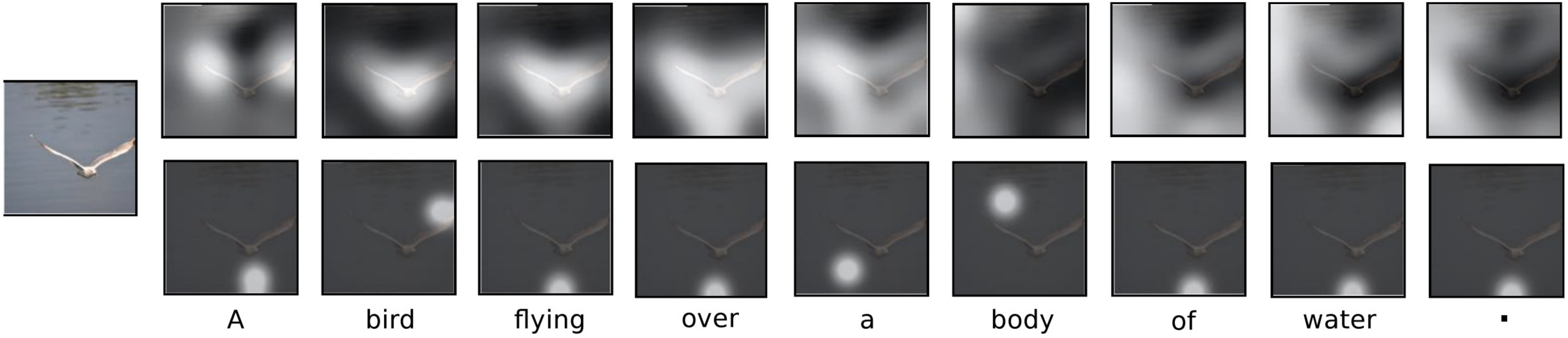
Attention on Images – Image Captioning

Hard attention *vs* Soft attention



Attention on Images – Image Captioning

Hard attention *vs* Soft attention



Attention on Images – Image Paragraph Generation

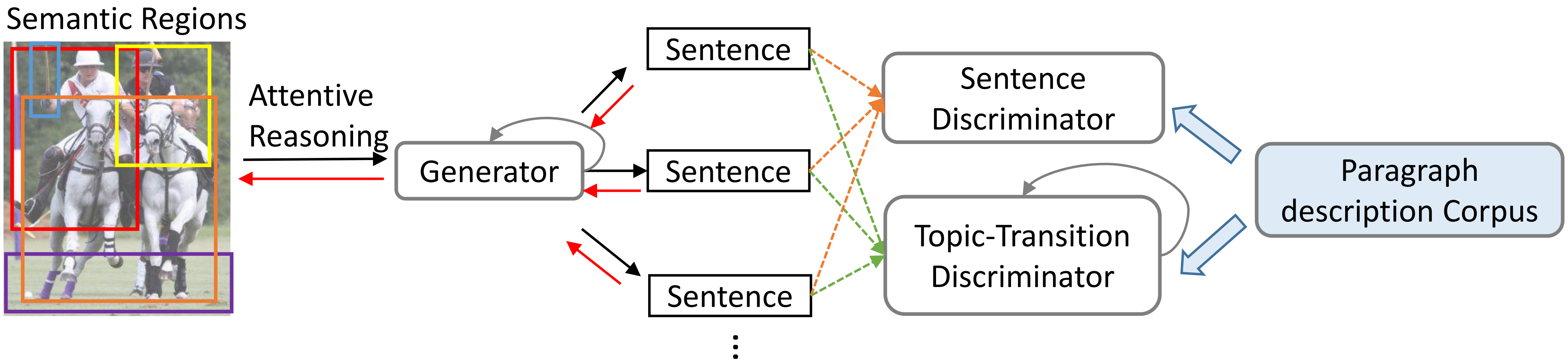
- Generate a long paragraph to describe an image
  - Long-term visual and language reasoning
  - Contentful descriptions -- ground sentences on visual features

This picture is taken for three baseball players on a field. The man on the left is wearing a blue baseball cap. The man has a red shirt and white pants. The man in the middle is in a wheelchair and holding a baseball bat. Two men are bending down behind a fence. There are words band on the fence.

A tennis player is attempting to hit the tennis ball with his left foot hand. He is holding a tennis racket. He is wearing a white shirt and white shorts. He has his right arm extended up. There is a crowd of people watching the game. A man is sitting on the chair.

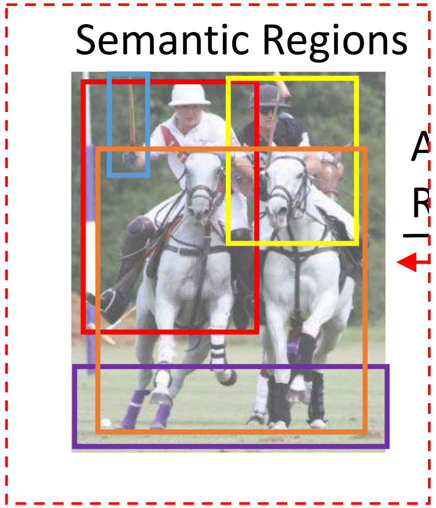
A couple of zebra are standing next to each other on dirt ground near rocks. There are trees behind the zebras. There is a large log on the ground in front of the zebra. There is a large rock formation to the left of the zebra. There is a small hill near a small pond and a wooden log. There are green leaves on the tree.

Attention on Images – Image Paragraph Generation

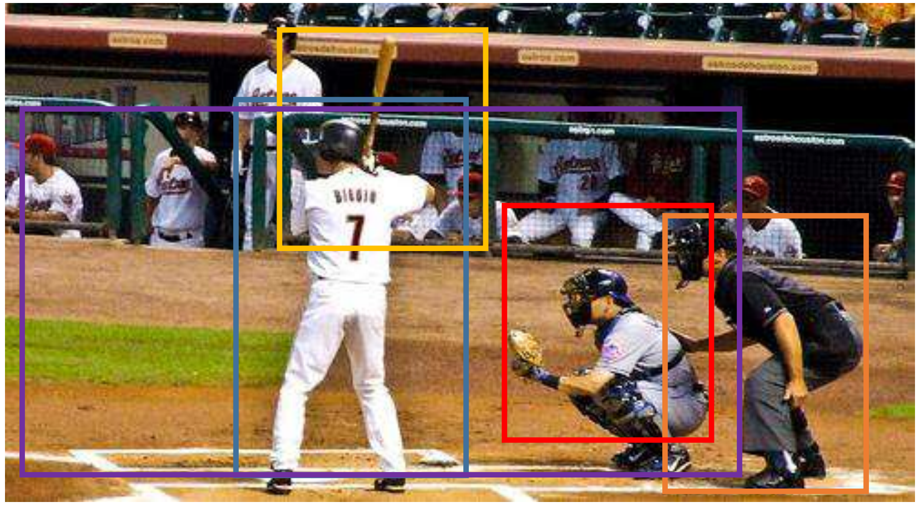


[Recurrent Topic-Transition GAN for Visual Paragraph Generation. Liang et al. 2017]

Attention on Images – Image Paragraph Generation



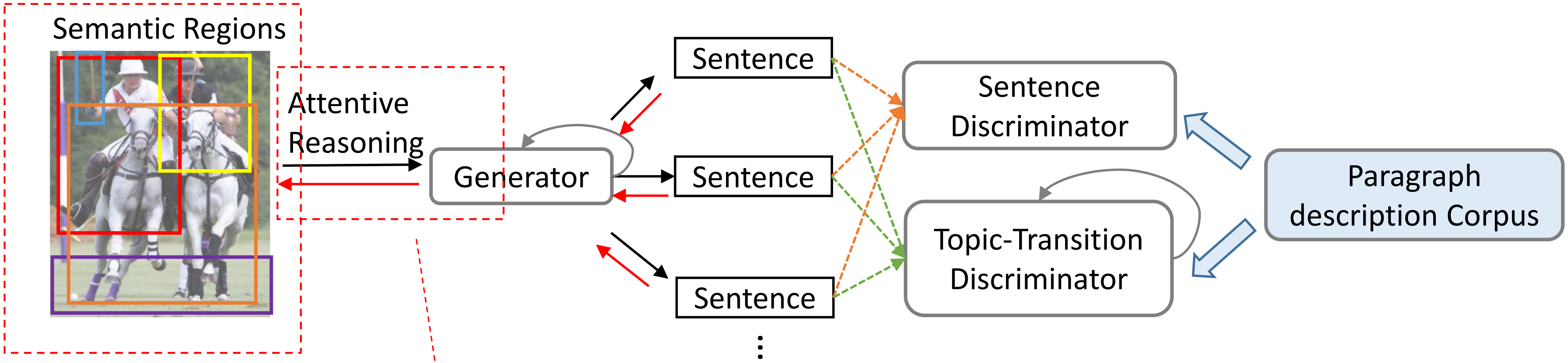
Semantic region detection & captioning



Local Phrases

- people playing baseball
- a man wearing white shirt and pants
- man holding a baseball bat
- person wearing a helmet in the field
- a man bending over

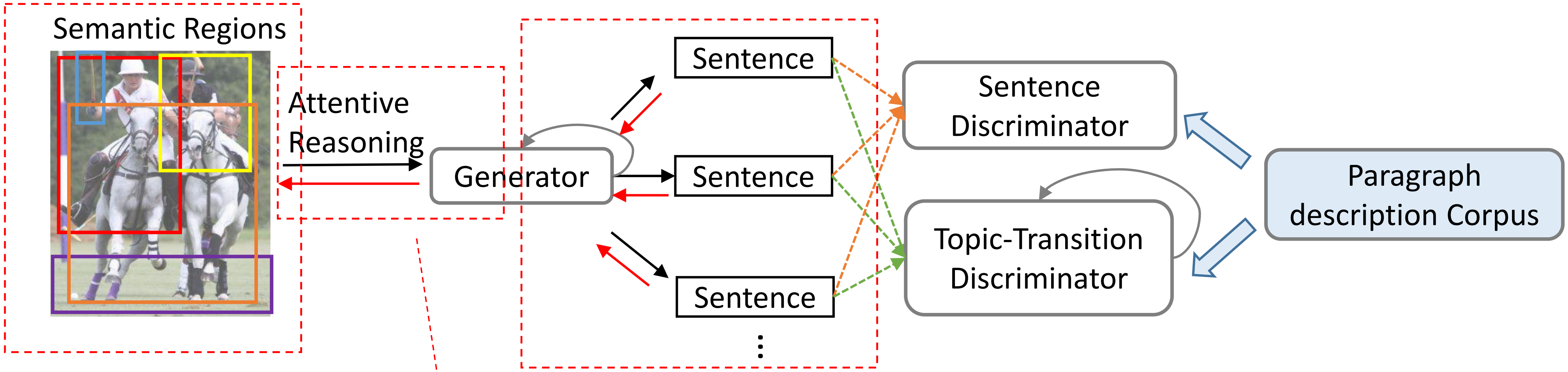
Attention on Images – Image Paragraph Generation



Semantic region
detection & captioning

Attention on both visual
regions and text
phrases

Attention on Images – Image Paragraph Generation

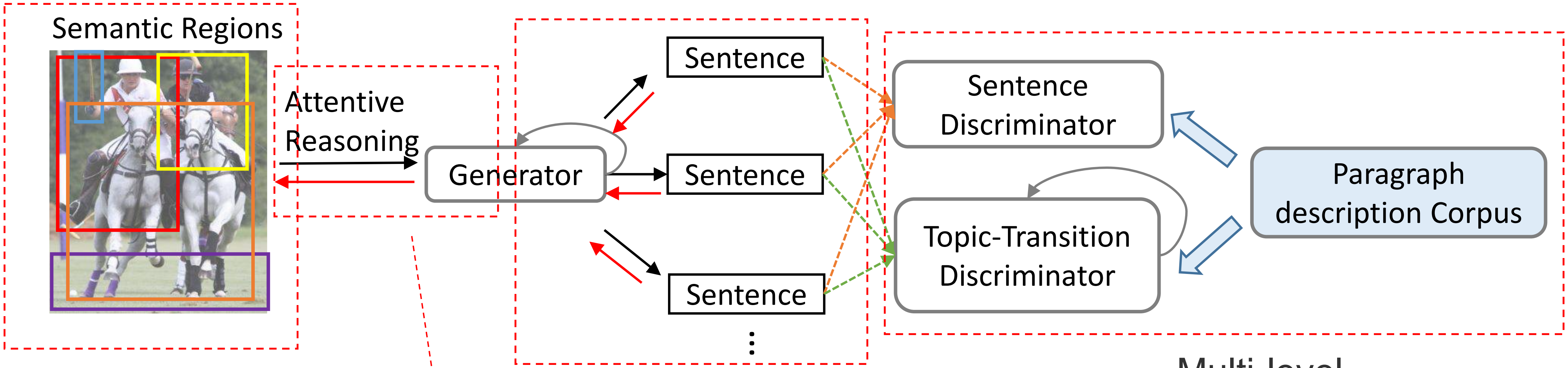


Semantic region detection & captioning

Hierarchical text generation

Attention on both visual regions and text phrases

Attention on Images – Image Paragraph Generation



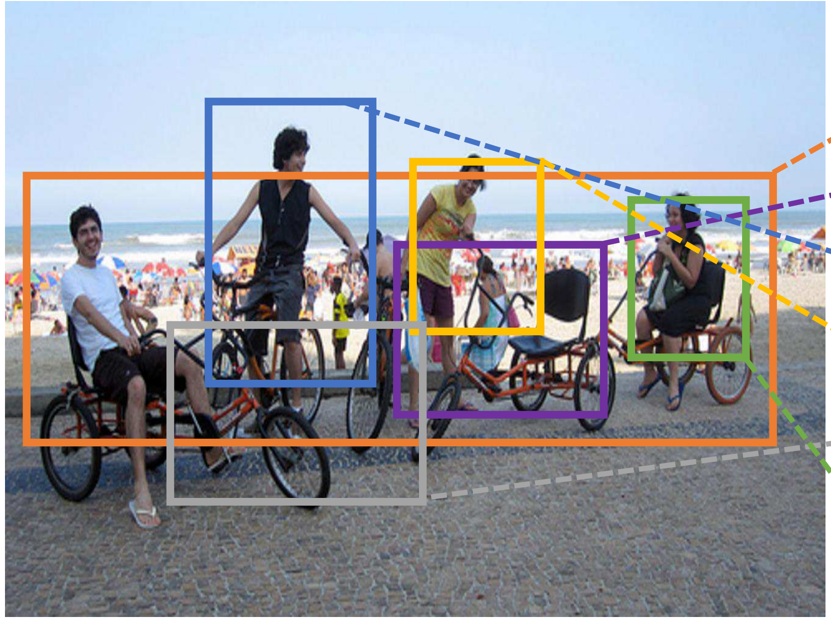
Semantic region detection & captioning

Hierarchical text generation

Multi-level adversarial learning

Attention on both visual regions and text phrases

Attention on Images – Image Paragraph Generation



1) people riding a bike
2) a bicycle parked on the sidewalk
3) man wearing a black shirt
4) a woman wearing a yellow shirt
5) a red and black bike
6) a woman wearing a shirt

Paragraph: *A group of people are riding bikes. There are two people riding bikes parked on the sidewalk. He is wearing a black shirt and jeans. A woman is wearing a short sleeve yellow shirt and shorts. There are many other people on the red and black bikes. A woman wearing a shirt is riding a bicycle.*

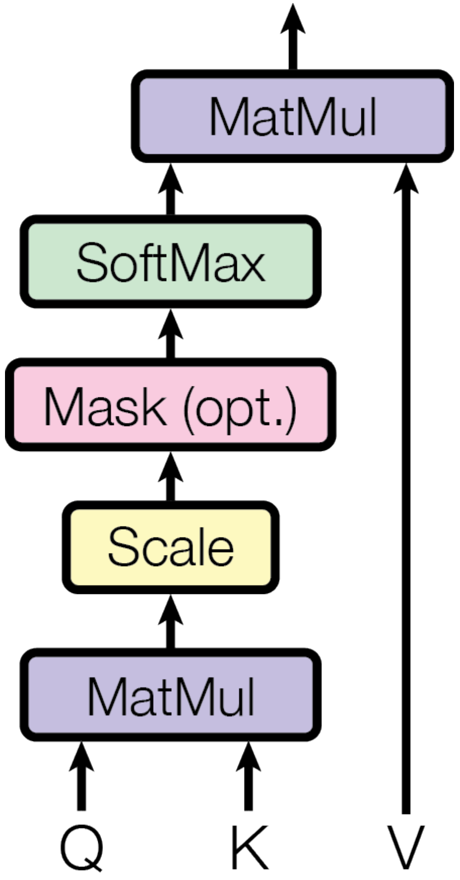
Outline

- Recurrent Networks (RNNs)
  - Long-range dependency, vanishing gradients
  - LSTM
  - RNNs in different forms
- Attention Mechanisms
  - (Query, Key, Value)
  - Attention on Text and Images
- Transformers: Multi-head Attention

Transformers – Multi-head (Self-)Attention

- State-of-the-art Results by Transformers
  - [Vaswani et al., 2017] Attention Is All You Need
    - Machine Translation
  - [Devlin et al., 2018] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
    - Pre-trained Text Representation
  - [Radford et al., 2019] Language Models are Unsupervised Multitask Learners
    - Language Models

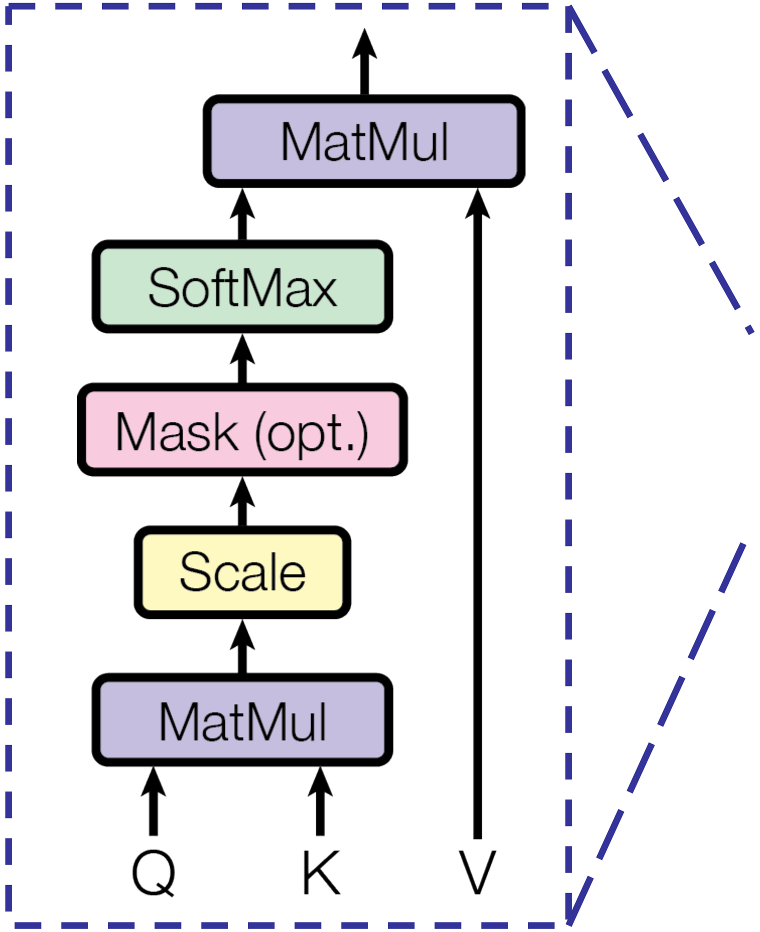
Multi-head Attention



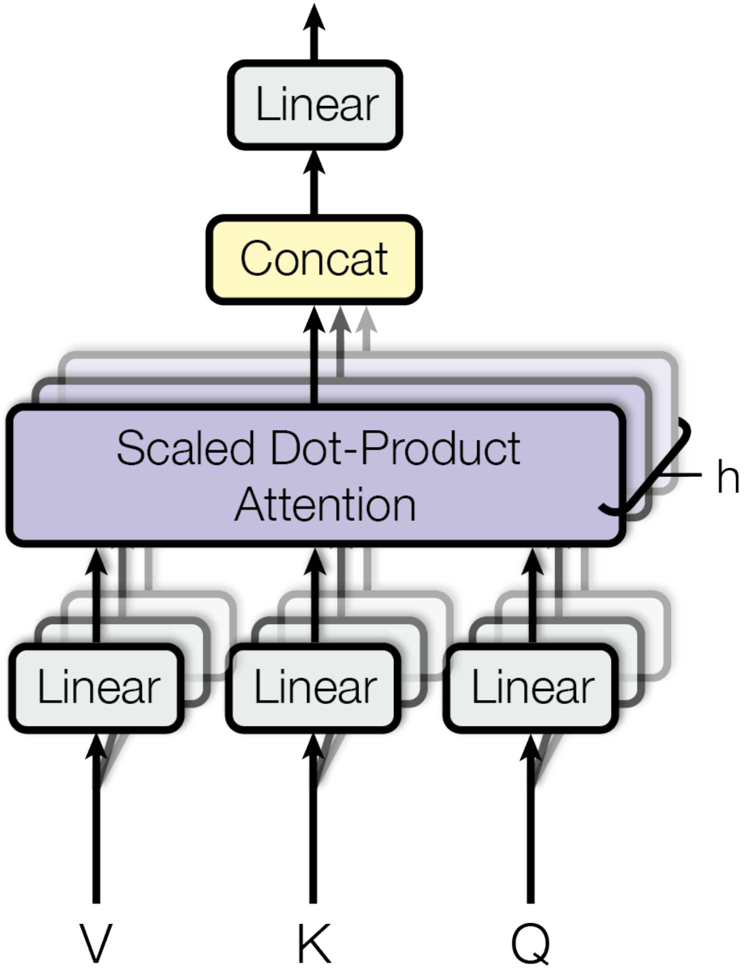
Scaled Dot-Product Attention

Image source: [Vaswani, et al., 2017]

Multi-head Attention



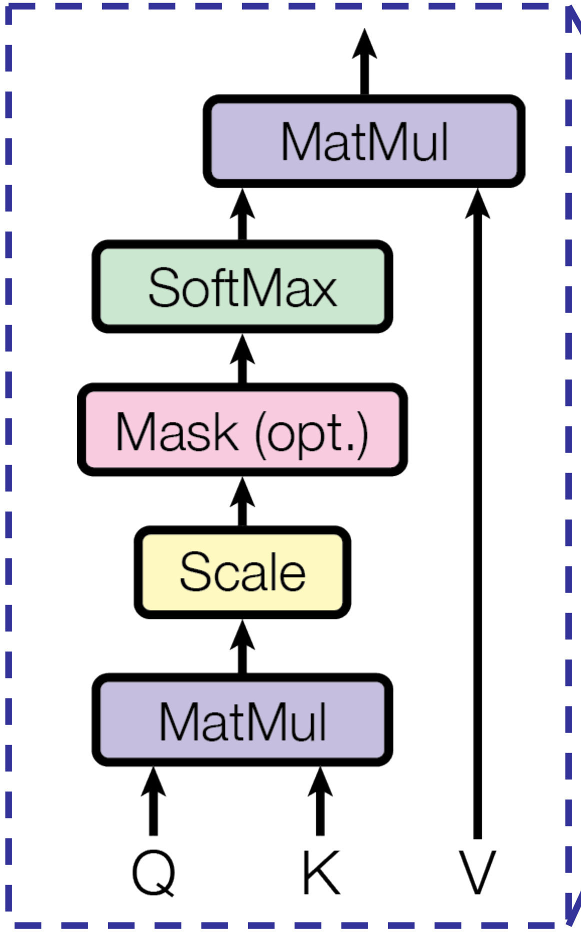
Scaled Dot-Product Attention



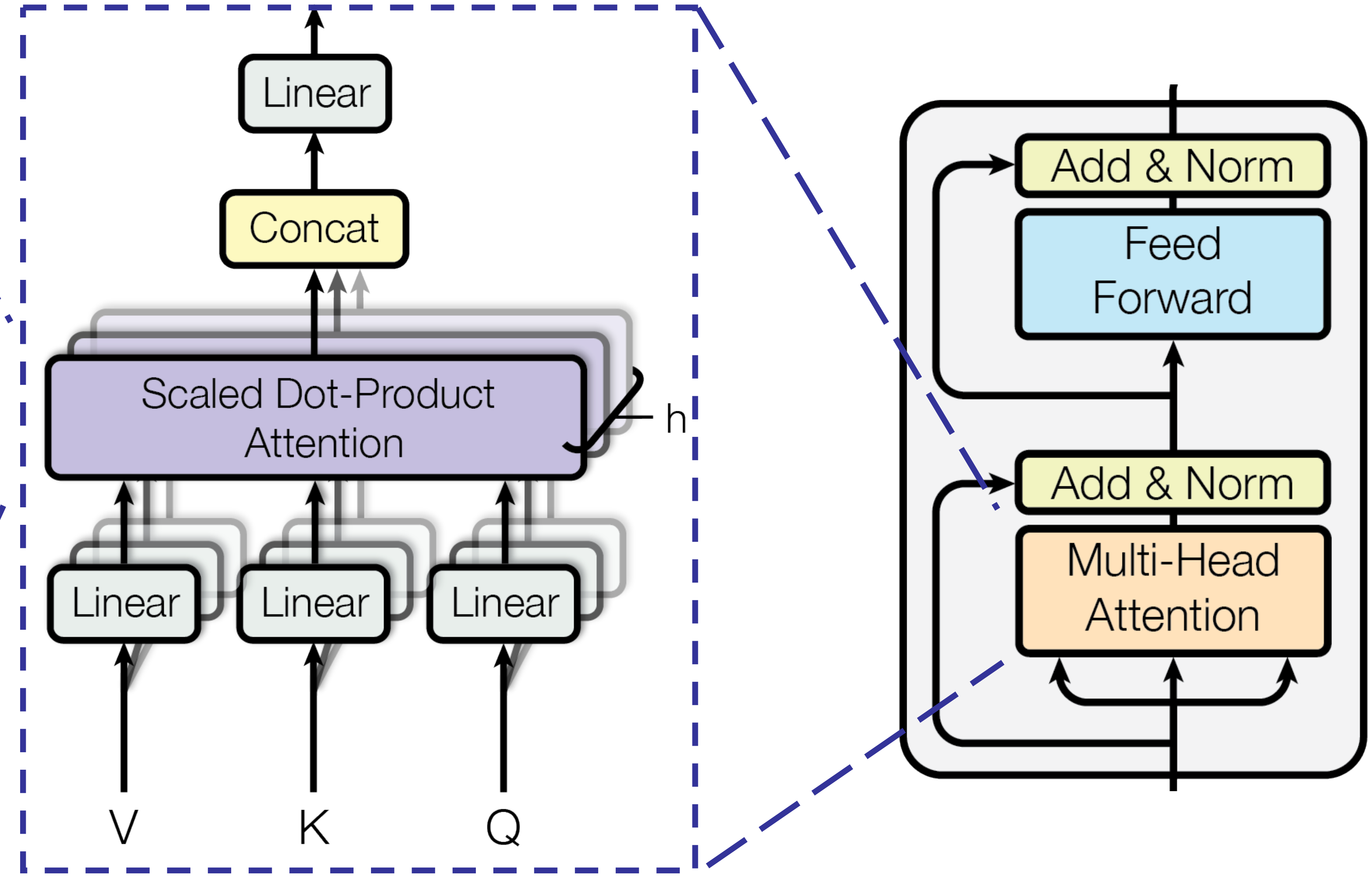
Multi-head Attention

Image source: Vaswani, et al., 2017

Multi-head Attention



Scaled Dot-Product Attention

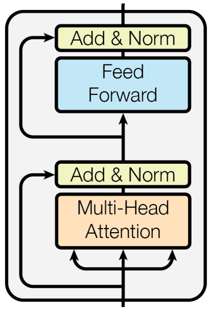


Multi-head Attention

Image source: Vaswani, et al., 2017

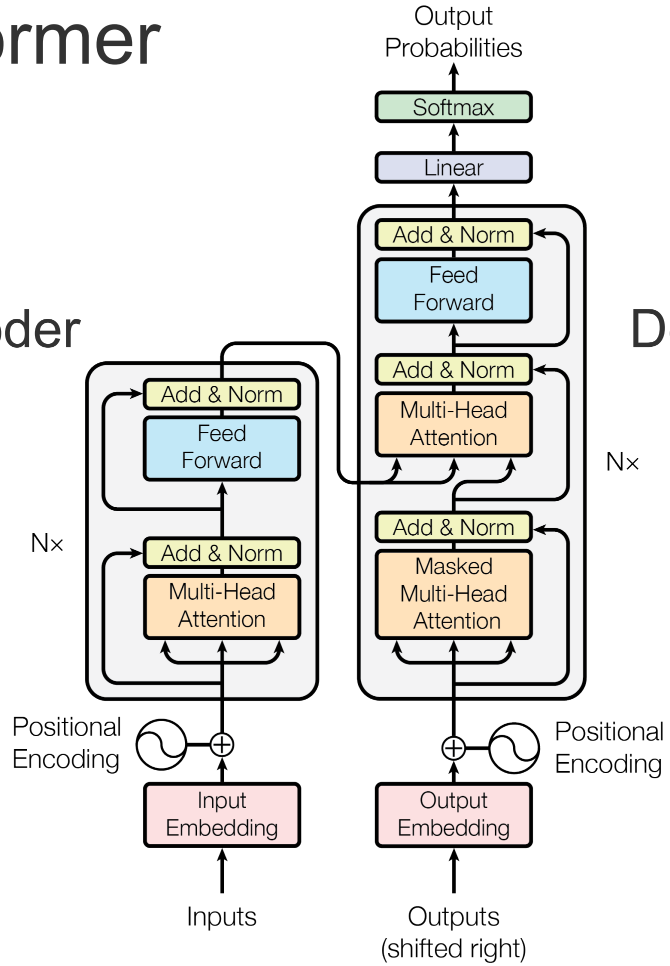
Multi-head Attention in Encoders and Decoders

Transformer



Encoder

Decoder



Multi-head Attention in Encoders and Decoders

Transformer

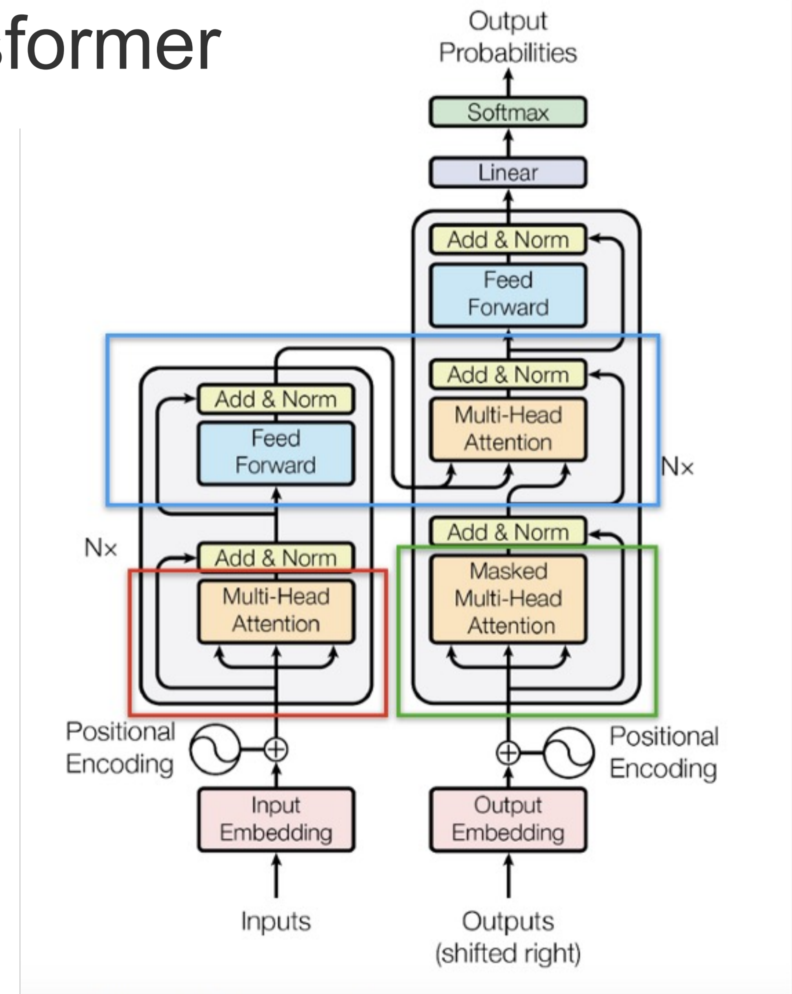


Figure 1: The Transformer - model architecture.

encoder self attention

1. Multi-head Attention
2. Query=Key=Value

decoder self attention

1. Masked Multi-head Attention
2. Query=Key=Value

encoder-decoder attention

1. Multi-head Attention
2. Encoder Self attention=Key=Value
3. Decoder Self attention=Query

Image source: Bgg

Questions?