Attention mechanisms in deep learning

Deep Learning for Human Brain Mapping, June 2020

Adriana Romero

Outline

- Motivation and overview
- Attention mechanisms
- Applications of attention
  - Attention in neural machine translation (RNN coupled with attention, the Transformer)
  - Attention in vision (image captioning, image-to-set prediction, image-to-recipe generation)
  - Attention in graphs (graph attention networks)
- Wrap Up

Motivation and overview

Motivation (1)

Images from https://unsplash.com/

Motivation (2)

John arrived early at the train station and waited until 2pm for the .....

Rachel was at home when the doorbell rang, she opened the ...

Motivation (2)

John arrived early at the train station and waited until 2pm for the train.

Rachel was at home when the doorbell rang, she opened the ...

Motivation (2)

John arrived early at the train station and waited until 2pm for the train.

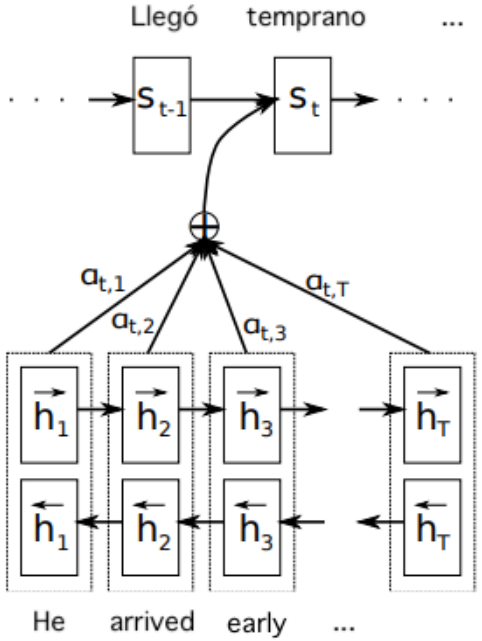
Rachel was at home when the doorbell rang, she opened the door.

Attention overview (1)

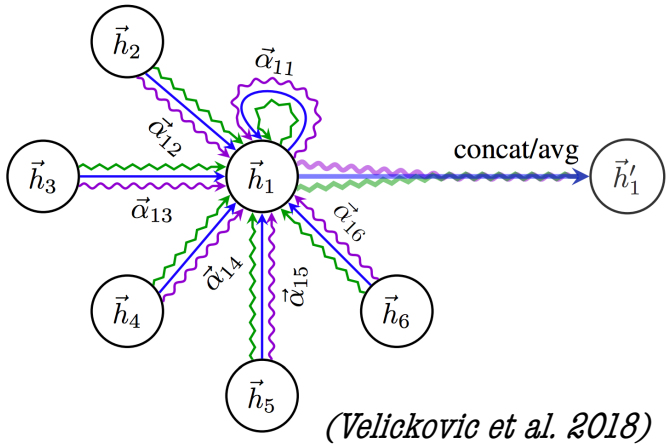
Attention mechanisms make use of the observation that different parts of data may have different significance, allowing us to concentrate on a subset of information and to select the most pertinent piece of information.

Attention overview (2)

(Xu et al. 2015)



(Bahdanau et al. 2015)



(Velickovic et al. 2018)

(Salvador et al. 2019)

Title Strawberry pie

Ingredients

- strawberries
- sugar
- flour
- butter

Instructions

1. Preheat oven to 350 degrees.
2. Combine butter, sugar , and flour in mixing bowl.
3. Cut in strawberries and set aside.

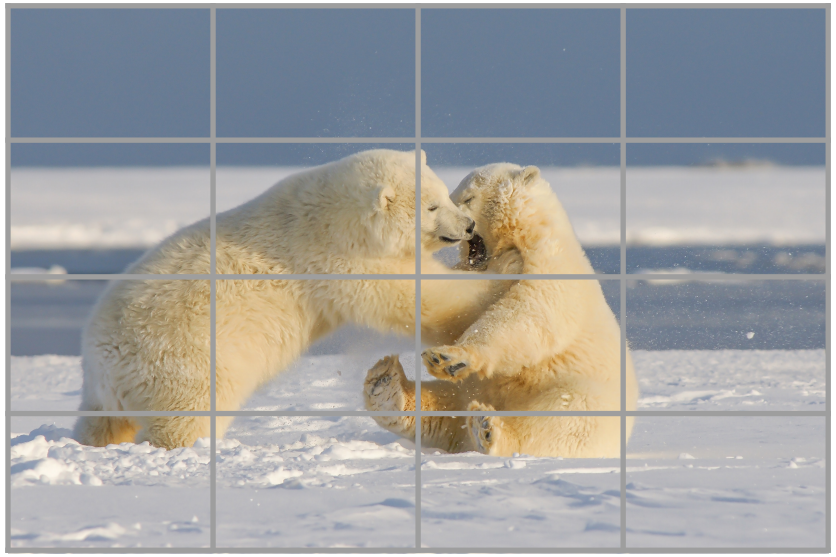
...

Attention can be applied to many application domains.

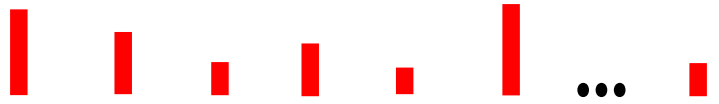
Attention mechanisms

Attention mechanisms (1)

Attention mechanisms (1)



Attention weights $\{\alpha_k\}_{k=1}^{N}$



"Polar ..."

Query, context

$\boldsymbol{h}_q$

Extract weights

Keys, annotations

$\{\boldsymbol{h}_k\}_{k=1}^{N}$

Image from https://unsplash.com/

Attention mechanisms (2)

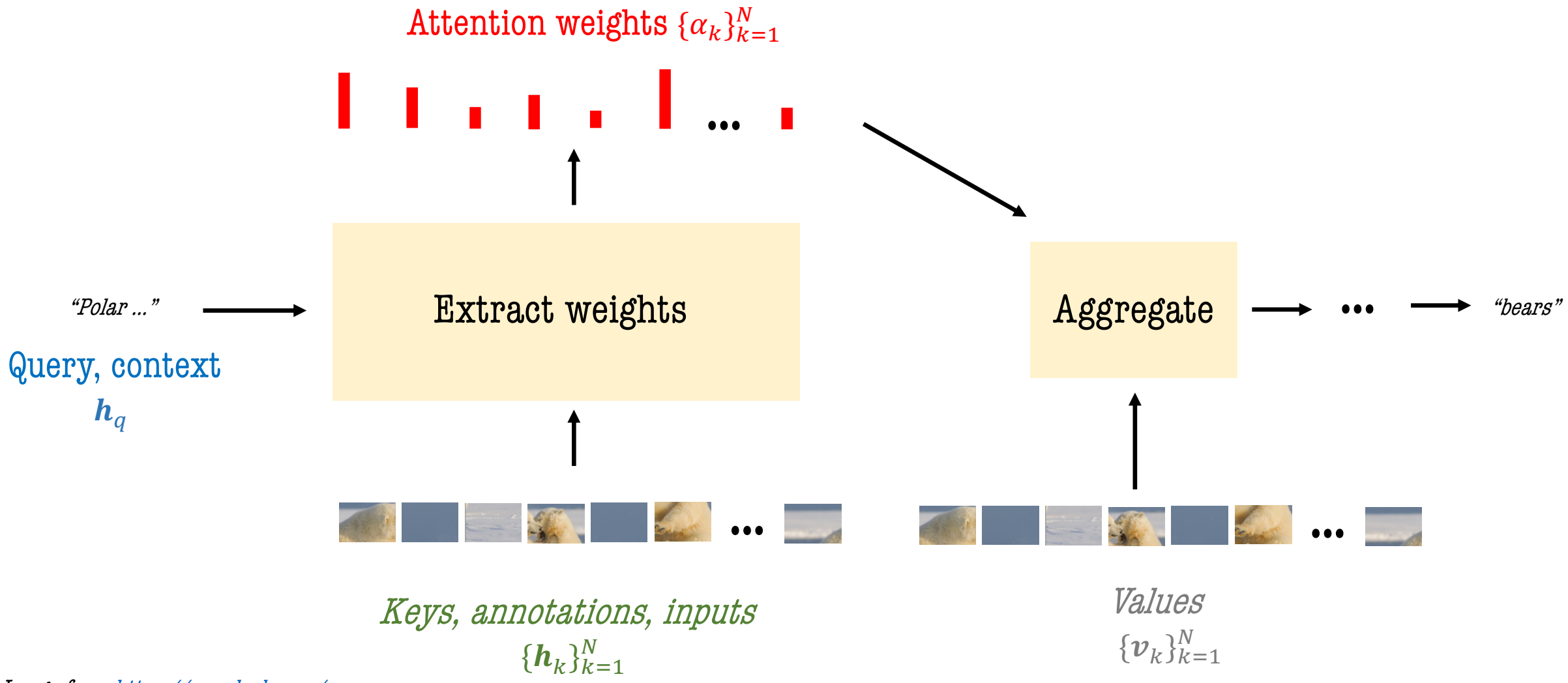


Image from https://unsplash.com/

Attention mechanisms (2)

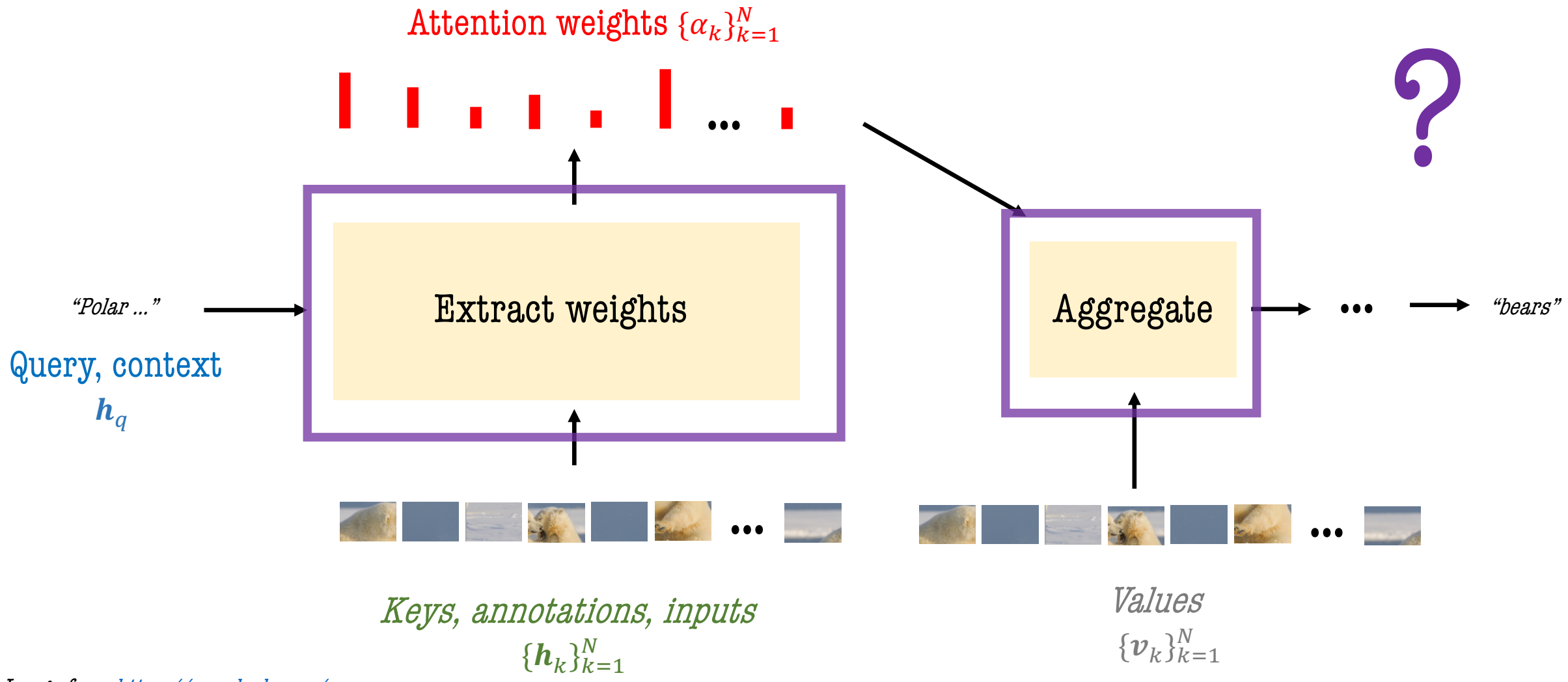
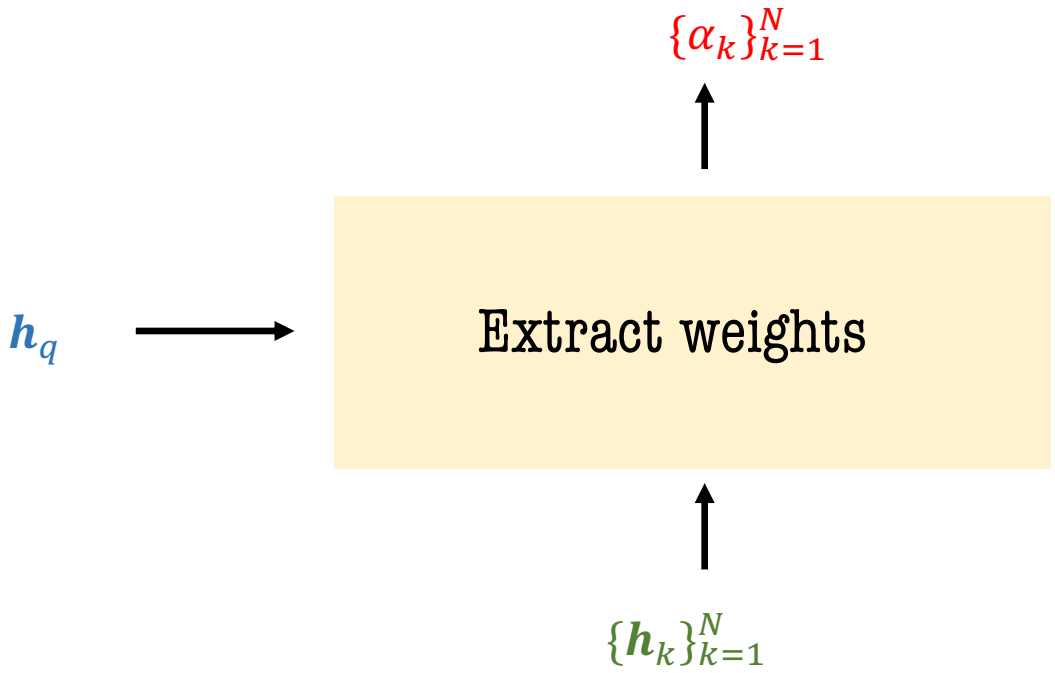


Image from https://unsplash.com/

Attention mechanisms (3)



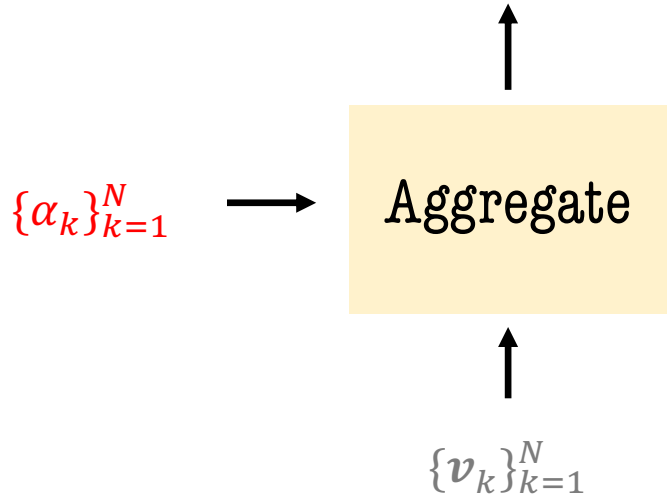
1. Combine key and query information

$$e_k = f_\theta(h_q, h_k)$$

2. Apply softmax to obtain attention weights α_k

$$\alpha_k = \frac{\exp(e_k)}{\sum_{k'=1}^{N} \exp(e_{k'})}$$

Attention mechanisms (4)



1. Combine key and query information

$$e_k = f_\theta(h_q, h_k)$$

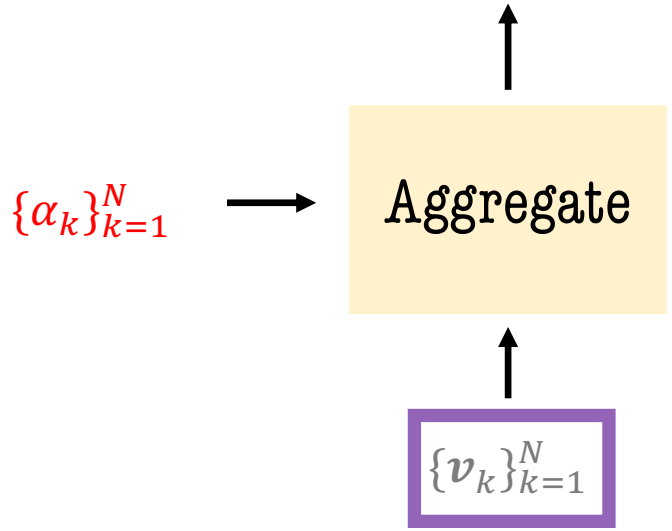
2. Apply softmax to obtain attention weights α_k

$$\alpha_k = \frac{\exp(e_k)}{\sum_{k'=1}^{N} \exp(e_{k'})}$$

3. Apply attention weights to values and aggregate

$$\sum_k \alpha_k v_k$$

Attention mechanisms (4)



1. Combine key and query information

$$e_k = f_\theta(h_q, h_k)$$

2. Apply softmax to obtain attention weights α_k

$$\alpha_k = \frac{\exp(e_k)}{\sum_{k'=1}^{N} \exp(e_{k'})}$$

3. Apply attention weights to values and aggregate

$$\sum_k \alpha_k v_k$$

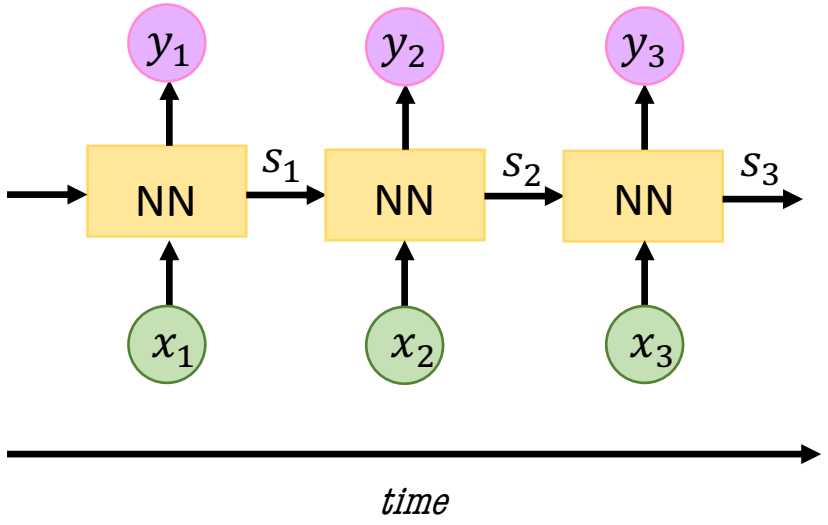
Applications of attention

Neural Machine Translation (1)

Neural Machine Translation (NMT): translating text from one language to another using neural networks.

"What are you doing today?" → *"¿Qué haces hoy?"*

- NMT long relied on encoder-decoder RNN and variants such as LSTM or GRU.

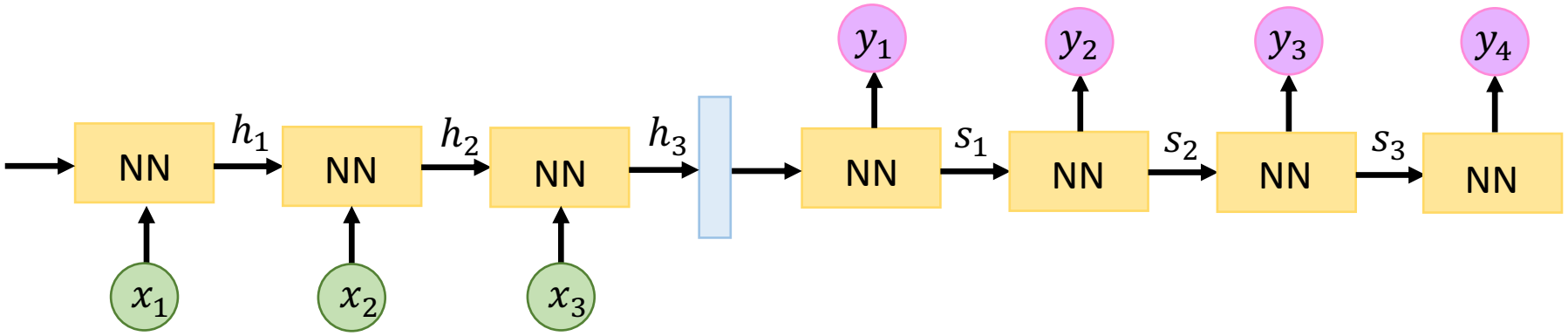


Neural Machine Translation (2)

Neural Machine Translation (NMT): translating text from one language to another using neural networks.

"What are you doing today?" ⟶ *"¿Qué haces hoy?"*

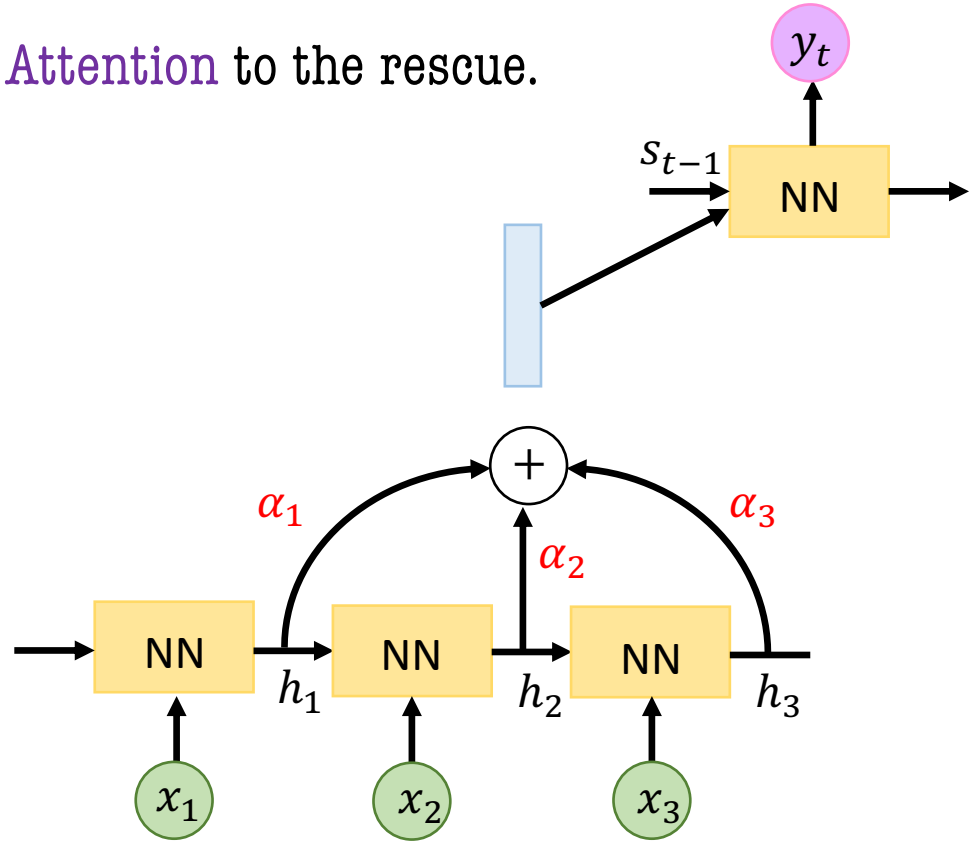
- NMT long relied on encoder-decoder RNN and variants such as LSTM or GRU.



Information bottleneck.

Attention in NMT (1)

- Attention to the rescue.



$$\boldsymbol{e}_k = f_\theta(\boldsymbol{s}_{t-1}, \boldsymbol{h}_k)$$

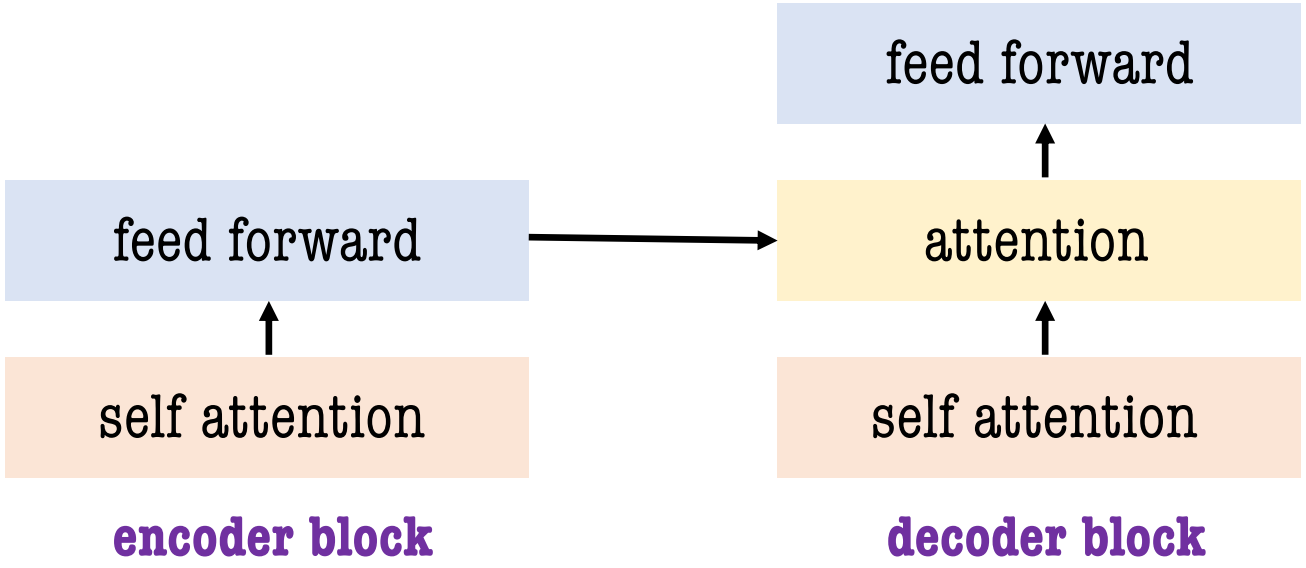
query: $\boldsymbol{s}_{t-1}$ (decoder hidden state at $t-1$)
keys: $\{\boldsymbol{h}_k\}_{k=1}^{N}$ (all encoder hidden states)
values = keys (all encoder hidden states)

f_θ is a feedforward neural network.

(Bahdanau et al. 2015)

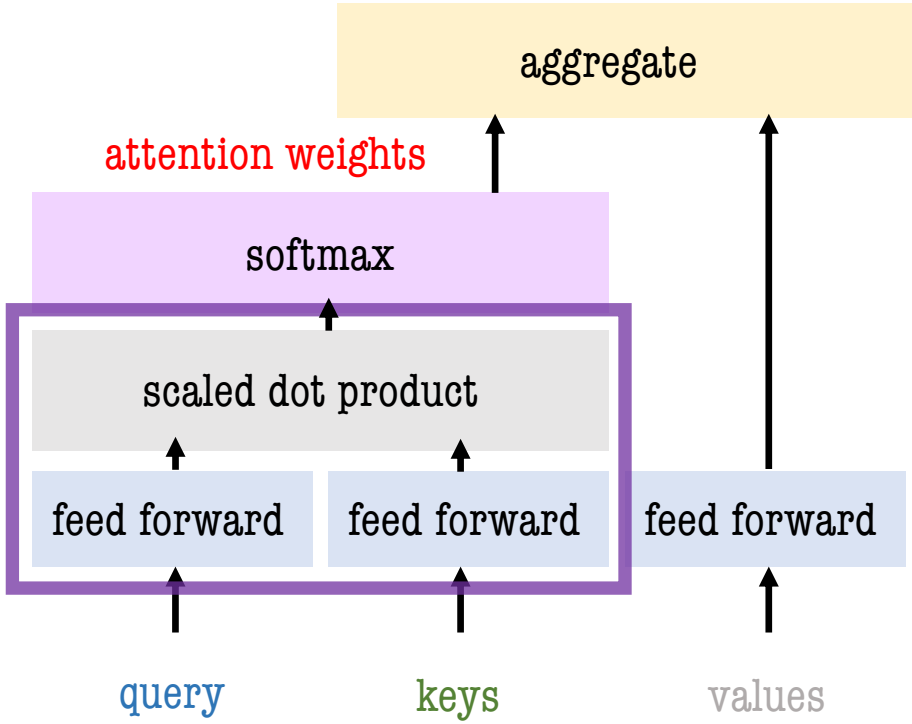
Attention in NMT (2)

The transformer: encoder-decoder model based solely on attention mechanisms, which is more parallelizable, and thus requires less time to train.



(Vaswani et al. 2017)

Attention in NMT (3)



self attention

query: word (embedding) from input/output sentence
keys: words (embeddings) from input/output sentence
values = keys

attention

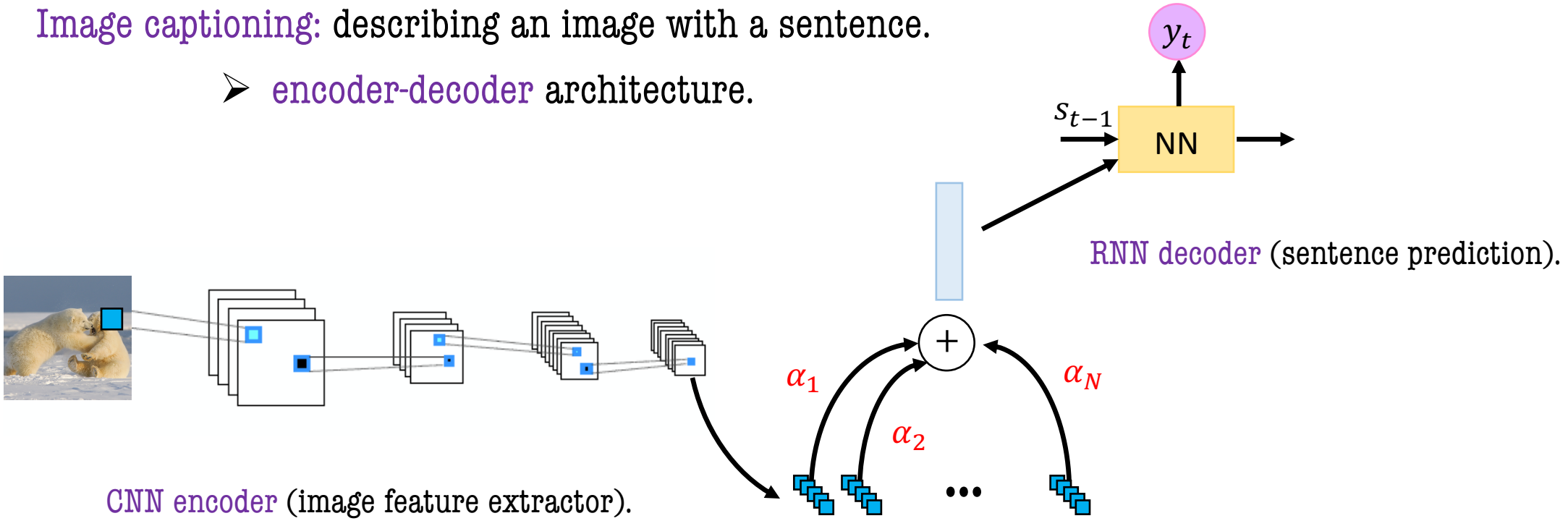
query : word (embedding) from output sentence
keys: words (embeddings) from input sentence
values = keys

(Vaswani et al. 2017)

Attention in vision (1)

Image captioning: describing an image with a sentence.

- encoder-decoder architecture.



query: $\boldsymbol{s}_{t-1}$ (decoder hidden state at $t-1$)
keys: $\{\boldsymbol{h}_k\}_{k=1}^{N}$ (image features)
values = keys (image features)

$$\boldsymbol{e}_k = f_\theta(\boldsymbol{s}_{t-1}, \boldsymbol{h}_k)$$

f_θ is a feedforward neural network.

Image from https://unsplash.com/

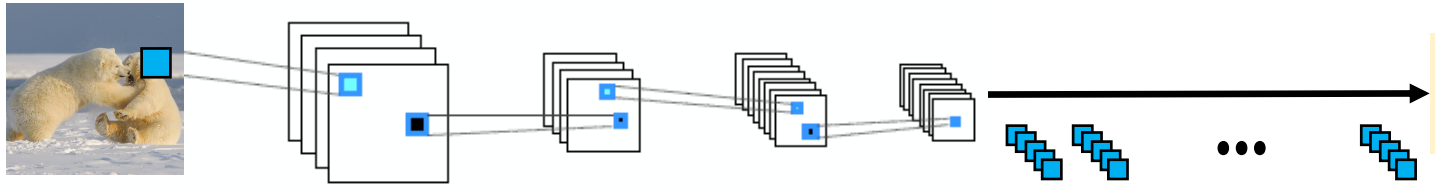
(Xu et al. 2015)

Attention in vision (2)

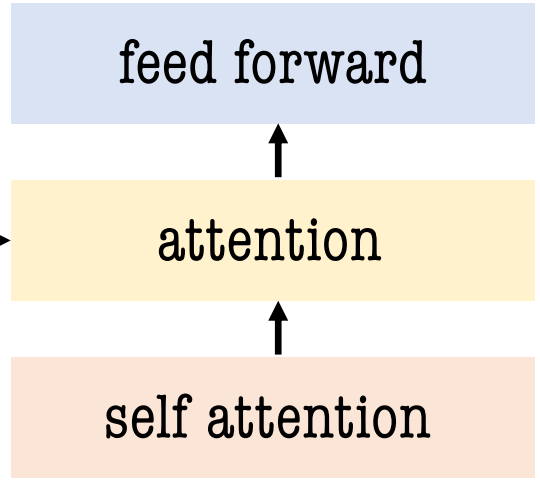
Image-to-set prediction (multi-label classification): describing an image with labels.

polar bear
snow
sky

- encoder-decoder architecture.



CNN encoder (image feature extractor).



query: predicted label
keys: image features
values = keys

query: predicted label
keys: predicted labels
values = keys

Transformer decoder

Image from https://unsplash.com/

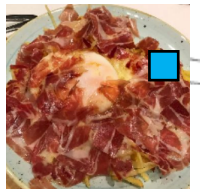
(Pineda et al. 2019)

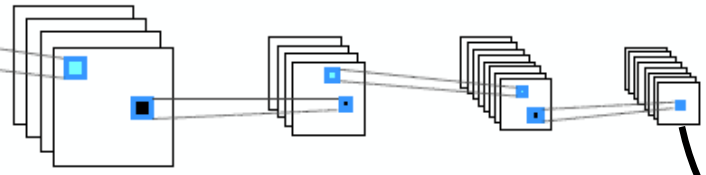
Attention in vision (3)

Image-to-recipe generation: writing recipe from food images.

Heat olive oil...
Add prosciutto...
Whisk eggs...

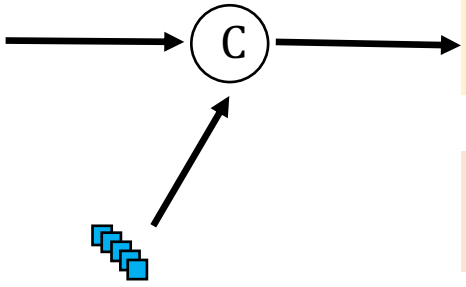
➢ encoder-decoder architecture.





ingredient prediction

Egg
Ham
...



CNN encoder (image feature extractor).

Transformer decoder

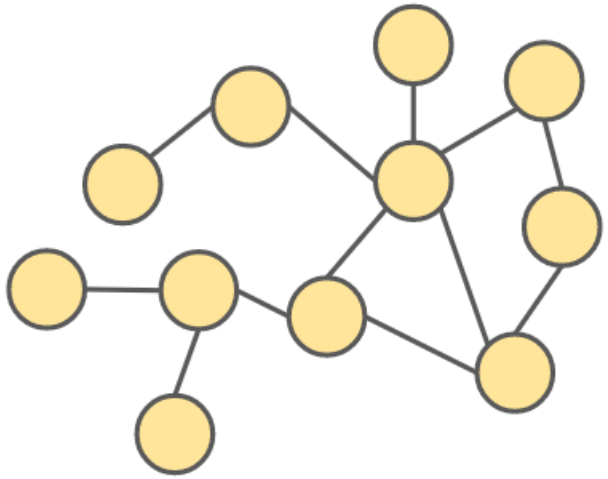
query: word (embedding) from output sentence
keys: words (embeddings) from output sentence
values = keys

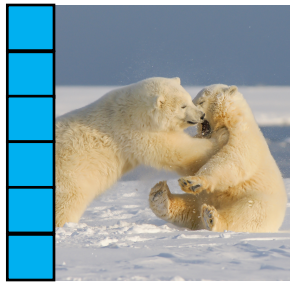
query : word (embedding) from output
keys : ingredients and image features
values = keys

(Salvador et al. 2019)

Attention in graphs

Graph Attention Networks: emulating convolutions on graphs by exploiting attention.





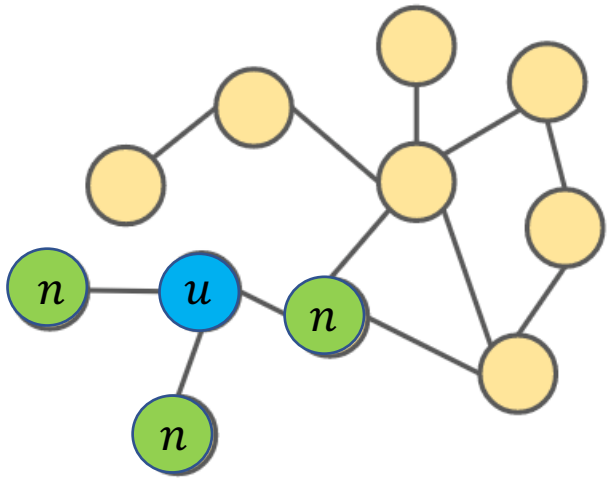
$$\sigma\left(\sum_i \alpha_i \boldsymbol{h}_i\right)$$

Image from https://unsplash.com/

(Velickovic et al. 2018)

Attention in graphs

Graph Attention Networks: emulating convolutions on graphs by exploiting attention.



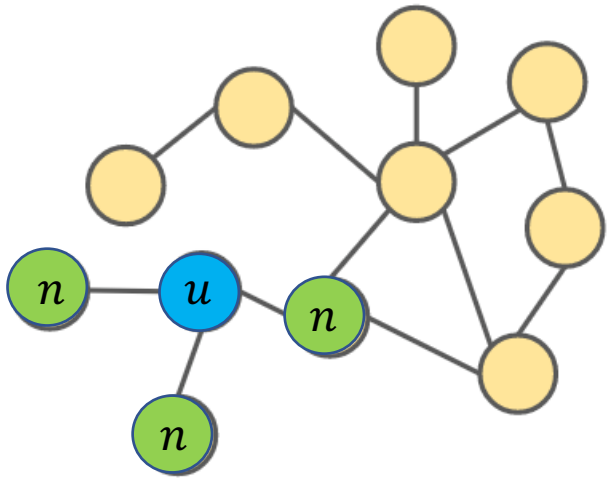
$$\sigma\left(\sum_{n \in \mathcal{N}(u) \cup \{u\}} \alpha_n \boldsymbol{h}_n\right)$$

- Attention to the rescue.

(Velickovic et al. 2018)

Attention in graphs

Graph Attention Networks: emulating convolutions on graphs by exploiting attention.



$$\sigma\left(\sum_{n \in \mathcal{N}(u) \cup \{u\}} \alpha_n \boldsymbol{h}_n\right)$$

- Attention to the rescue.

To obtain: α_n:

$$\boldsymbol{e}_k = f_\theta\left(\boldsymbol{h}_u, \boldsymbol{h}_n\right)$$

query: features of node u
keys: features of nodes $\mathcal{N}(u) \cup \{u\}$
values = keys

The same procedure is applied to all nodes.

(Velickovic et al. 2018)

Wrap Up

Wrap Up

Intro

- Motivating examples
- Attention overview

Attention mechanism

- Attention model
  - Query & keys fed to a scoring function
  - Scores passed through a softmax to obtain attention weights
- Attention weights and values combined in an aggregate step to compute the output of the attention mechanism

Attention Applications

- Neural Machine Translation
  - RNN coupled with attention
  - Transformer
- Vision
  - Image captioning
  - Image-to-set prediction
  - Image-to-recipe generation
- Graphs
  - Graph attention networks

Same principle, different definition of query/key/values and different scoring functions.

References

(Bahdanau et al., 2015) D. Bahdanau, K. Cho & Y. Bengio; **Neural Machine Translation by Jointly Learning to Align and Translate**, ICLR, 2015.
(Vaswani et al., 2017) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser & I. Polosukhin; **Attention is all you need**, NeurIPS, 2017.
(Xu et al., 2015) K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio; **Show, Attend and Tell: Neural Image Caption Generation with Visual Attention**, ICML, 2015.
(Pineda et al., 2019) L. Pineda, A. Salvador, M. Drozdzal, A. Romero; **Elucidating image-to-set prediction: An analysis of models, losses and datasets**, preprint arXiv:1904.05709, 2019.
(Salvador et al., 2019) A. Salvador, M. Drozdzal, X. Giro-i-Nieto, A. Romero; **Inverse Cooking: Recipe Generation from Food Images**, CVPR, 2019.
(Velickovic et al., 2018) P. Velickovic, G. Cucurull, A. Casanova, A. Romero. P. Lio & Y. Bengio; **Graph Attention Networks**, ICLR, 2018.

Attention mechanisms in deep learning

Deep Learning for Human Brain Mapping, June 2020

Adriana Romero