University of Glasgow | School of Computing Science

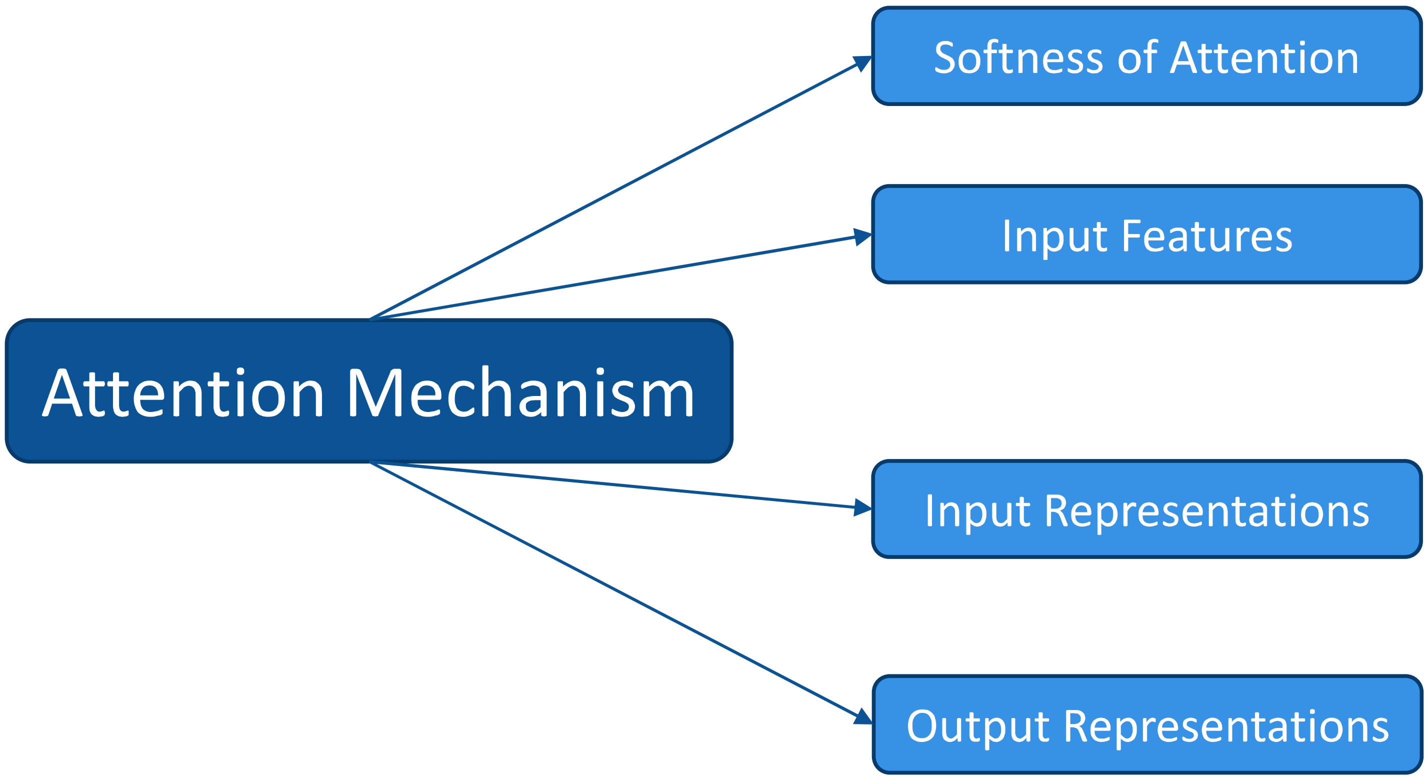
THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Taxonomy of Attention

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

WORLD CHANGING GLASGOW

Taxonomy of Attention



Softness of Attention
- Soft/Hard
- Global/Local

Input Features
- Item wise
- Location wise

Input Representations
- Distinctive
- Self-attention
- Co-attention
- hierarchical

Output Representations
- Single output
- Multi-head
- Multi-dimension

Niu et al. A review on the attention mechanism of deep learning, Neurocomputing, 2021

Hard vs Soft Attention

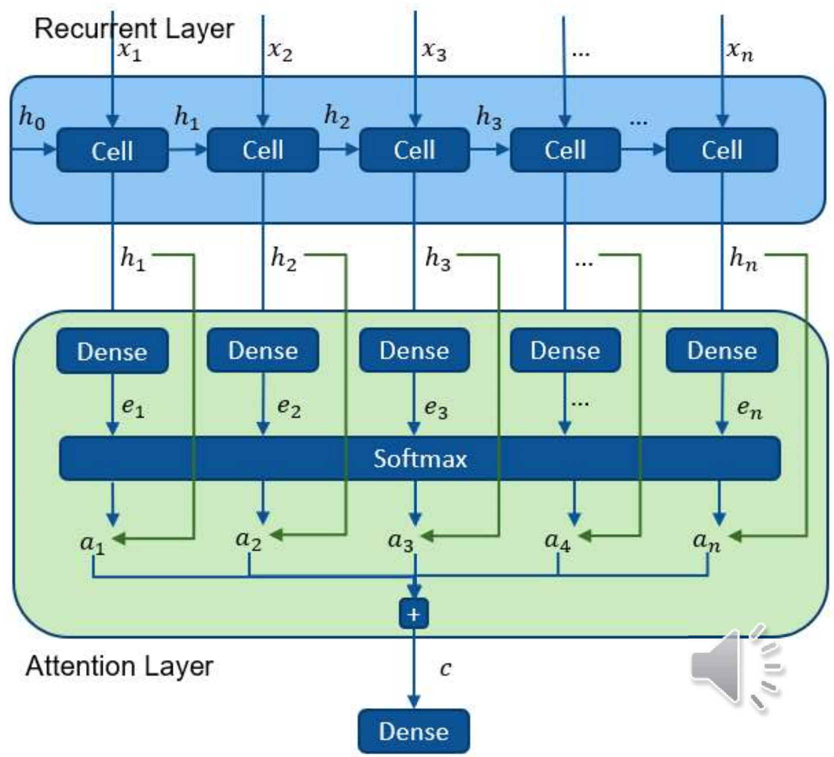
- Soft Attention
  - Attention score is used as weights in the weighted average context vector calculation
  - This is a **differentiable function**
  - The system is optimized by standard backpropagation

$$c = \sum_j \alpha_j h_j$$

Hard vs Soft Attention

- Soft Attention
  - Attention score is used as weights in the weighted average context vector calculation
  - This is a **differentiable function**
  - The system is optimized by standard backpropagation

$$c = \sum_j \alpha_j h_j$$



Hard vs Soft Attention

- Soft Attention
  - Attention score is used as weights in the weighted average context vector calculation
  - This is a **differentiable function**
  - The system is optimized by standard backpropagation

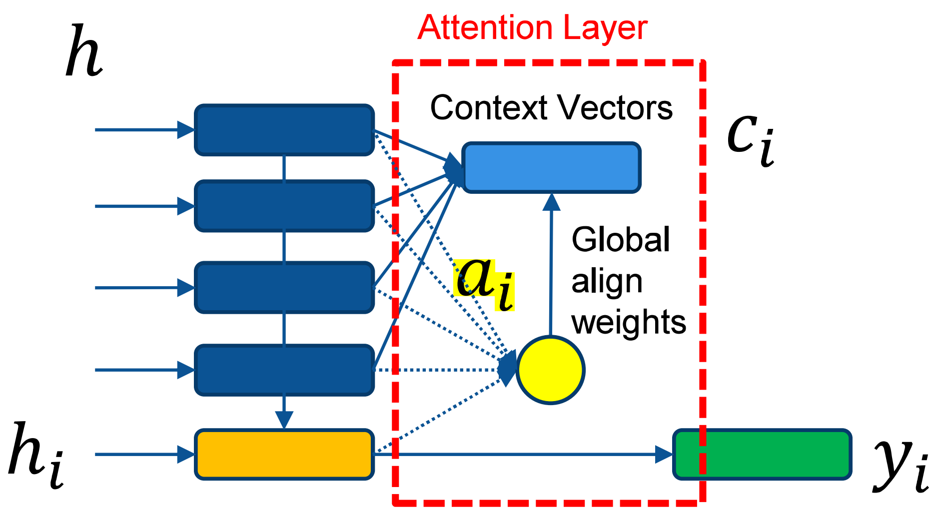
$$c = \sum_j \alpha_j h_j$$

- Hard Attention
  - The context vector is computed from stochastically sampled keys
  - It is **not differentiable**
  - Optimization cannot be performed with backpropagation (ie. reinforcement learning)

$$\tilde{a} \sim Multinoulli\ (\{\alpha_j\})$$

$$c = \sum_j \tilde{a}_j h_j$$

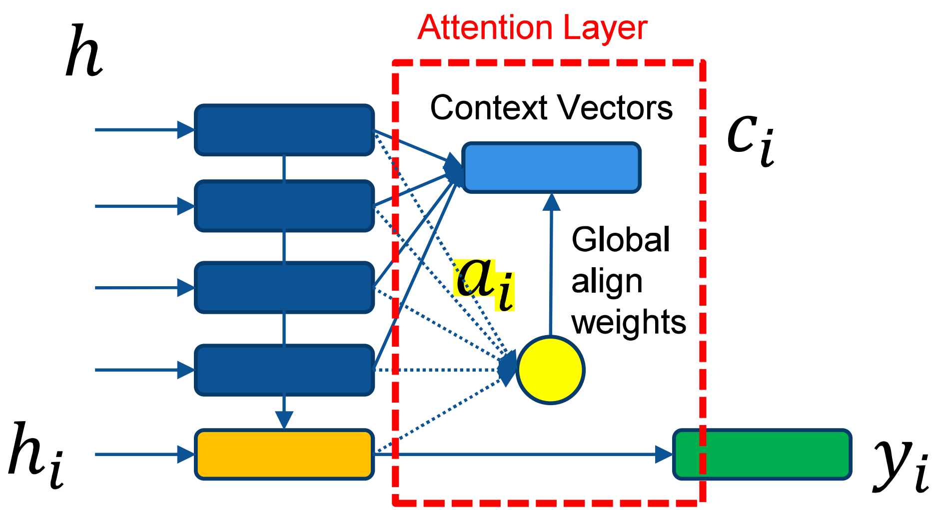
Global vs Local Attention



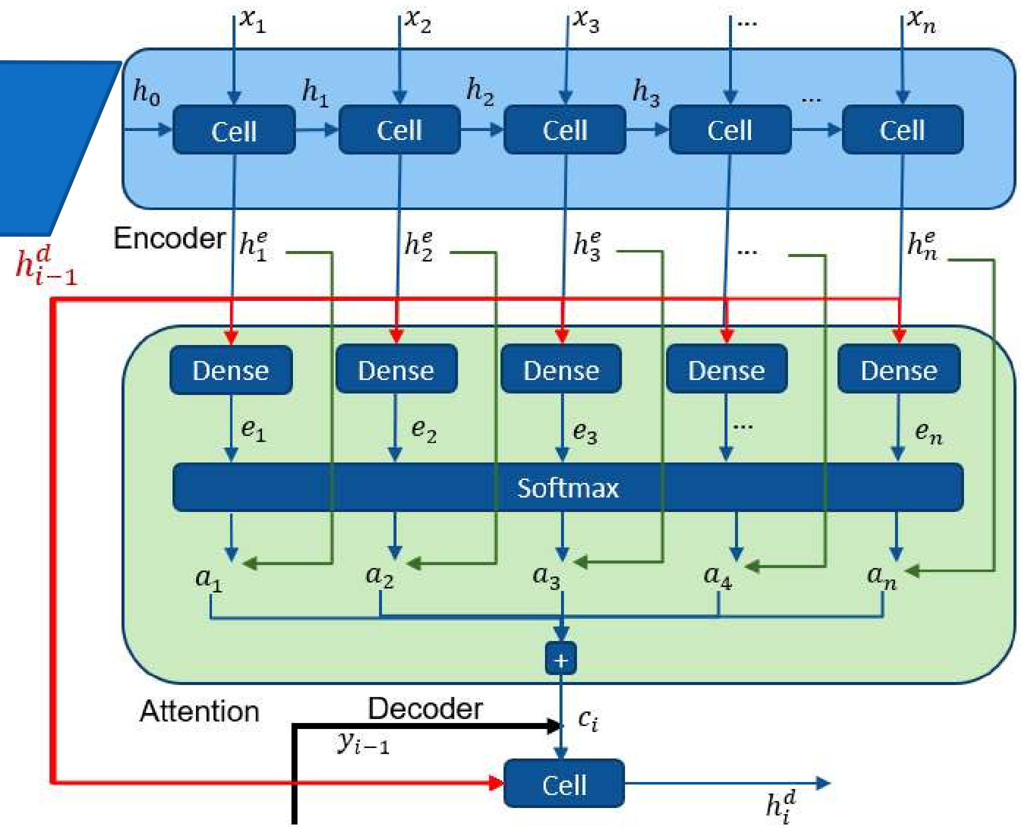
Global attention

- **Global attention** is like soft attention

Global vs Local Attention



Global attention



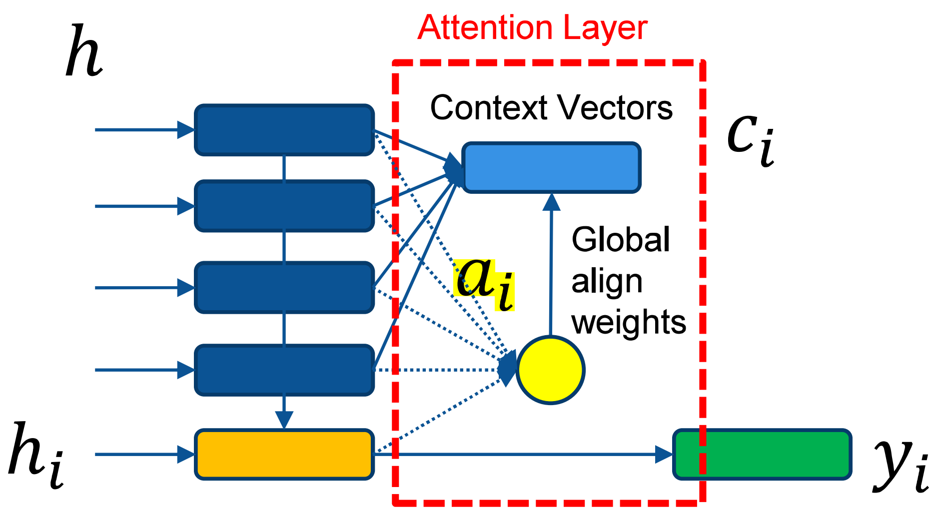
$$e_{ij} = a(h_j^e, h_i^d) = \tanh(W * h_j^e + U * h_{i-1}^d)$$

$$\alpha_{ij} = (softmax(e_i)_j) = \frac{\exp(e_{ij})}{\sum \exp(e_{ik})}$$

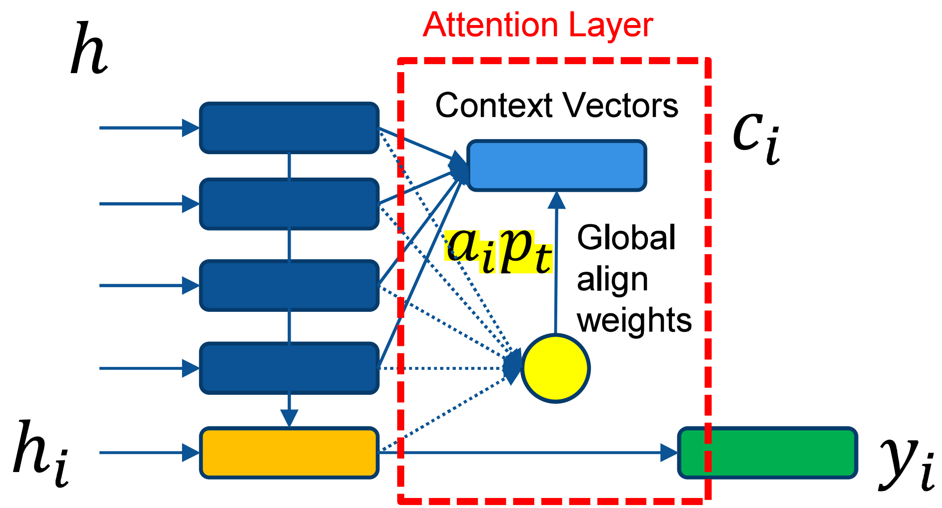
$$c_i = \sum_j \alpha_{ij} h_j^e$$

- **Global attention** is like soft attention

Global vs Local Attention

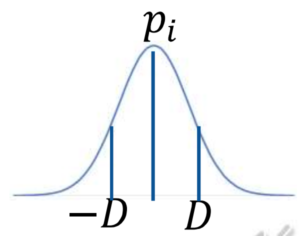


Global attention



Local attention

$$c_i = \sum_{j=p_i-D}^{p_i-D} \alpha_{ij} h_j^e$$



- **Global attention** is like soft attention
- **Local attention** is at the middle-ground between soft and hard attention

Forms of Input Features

- Item-wise if the input is a sequence of items
  - Each item is encoded separately
  - Combined with soft-attention estimates a weight for each item and subsequently it combines linearly
- Location-wise are suited for visual tasks
  - Accepts an entire feature map
  - Generates a transformed version through the attention module

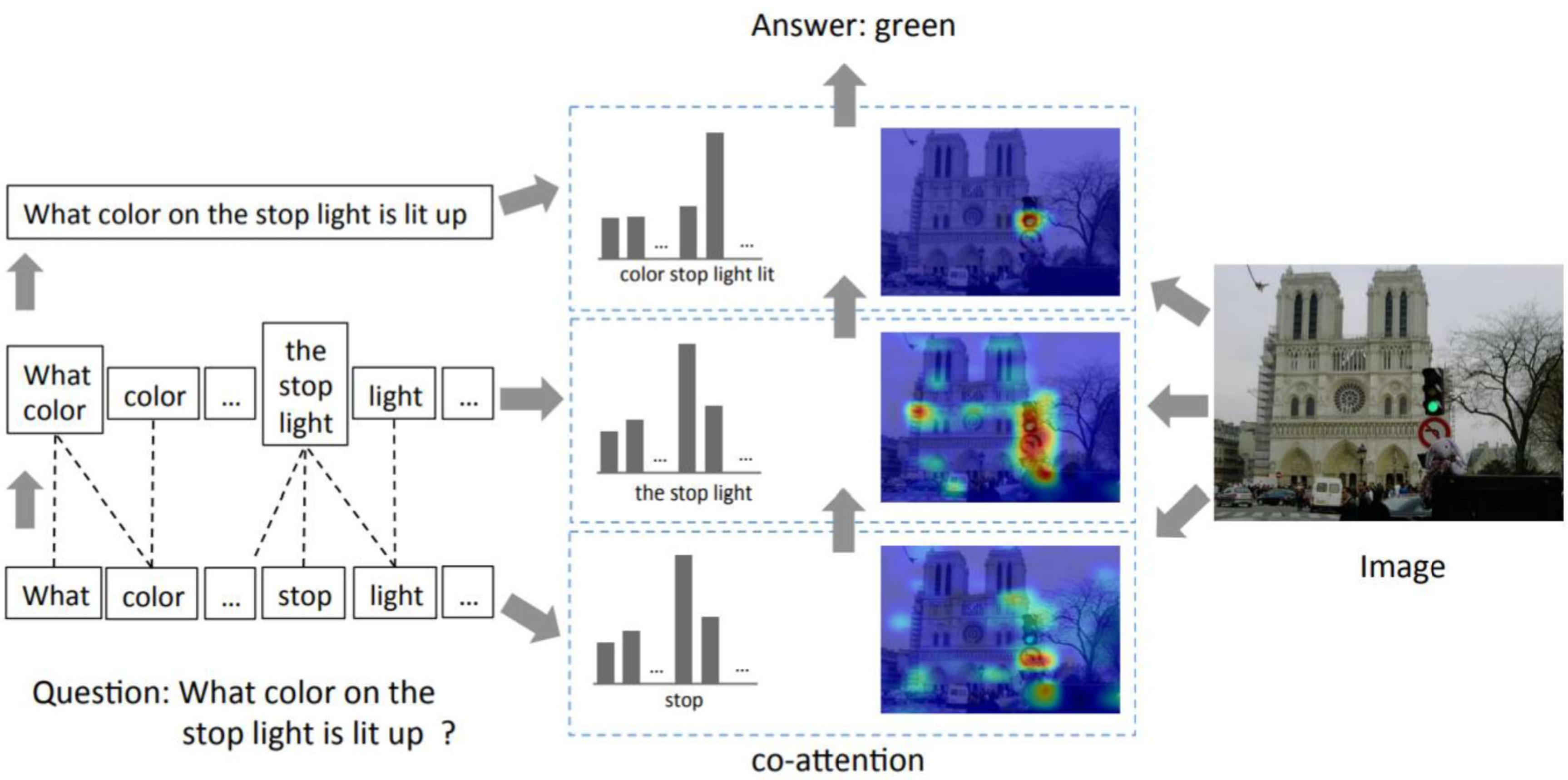
Input Representation

- Distinctive
  - Keys and queries belong to two independent sequences
- Self-Attention
  - Estimated based on the keys, without the need of queries
- Co-Attention
  - Jointly reason about multi-modal data, ie. Images and text in Q&A sessions
- Hierarchical
  - Attention estimated from different abstraction levels

Input Representation – Self Attention

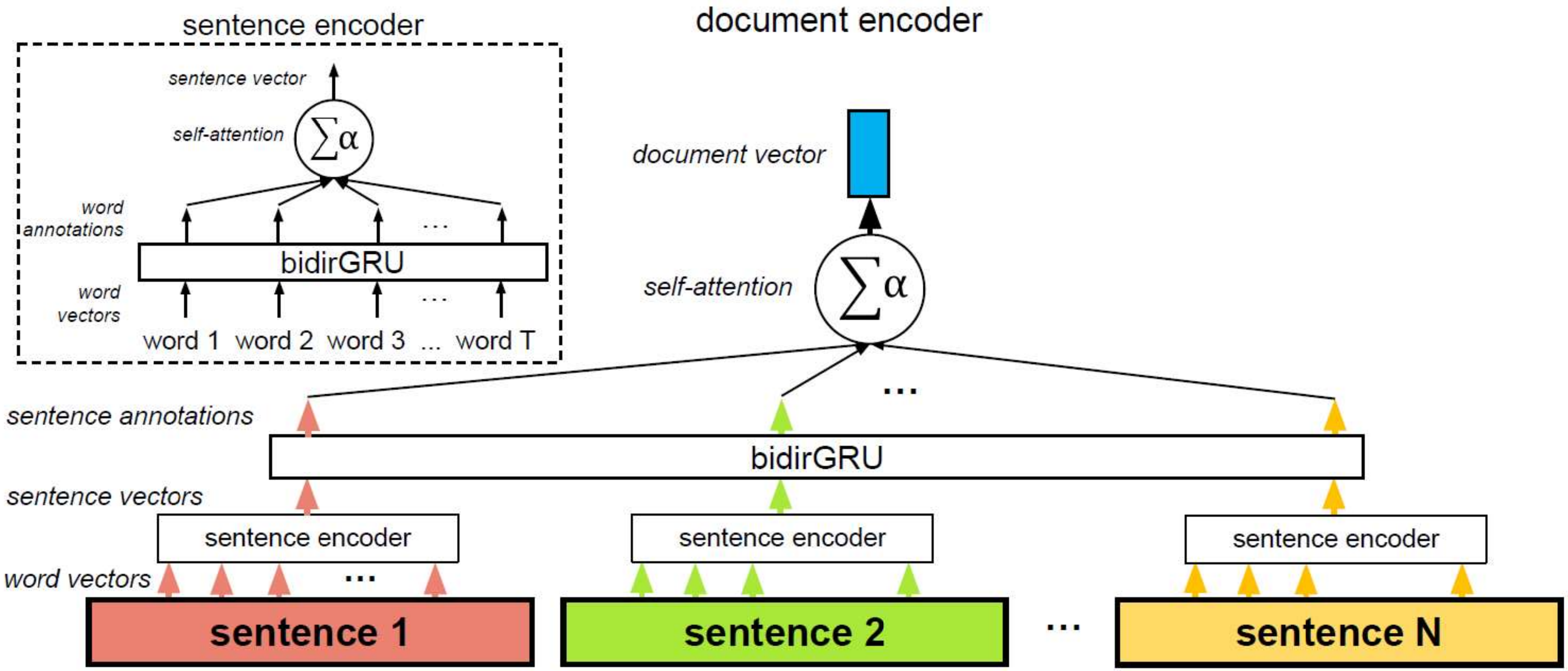
- Self-Attention
  - Estimated based on the keys, without the need of queries
  - It applies within a single layer without connecting two components
  - Several successful applications, ie. Transformers
  - It models dependencies between different parts of the input well

Input Representation: Co-Attention



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

Input Representation: Hierarchical Attention



Antoine J.-P. Tixier, Notes on Deep Learning for NLP, 2018

Output Representation

- Single Output
  - Single feature representation in each time step
  - Energy scores are presented as one vector at each time-step
- Multi-Head Output Attention
  - Linearly projects the input sequence to multiple channels
- Multi-Dimensional Output Attention
  - Calculates multiple attention distributions for the same data

Summary

- Attention mechanisms have been categorized in several different types
  - Soft or hard weights
  - Input features (item-wise or location-wise)
  - Input representation (self-attention, co-attention, hierarchical attention)
  - Output representation (single head, multi-head)

References

- Niu et al. A review on the attention mechanism of deep learning, Neurocomputing, 2021
- Foster, Generative Deep Learning – Teaching Machines to Paint, Write, Compose and Play, O’Reilly, 2019
- Lu et al. Hierarchical Question-Image Co-Attention for Visual Question Answering, NIPS, 2016
- Antoine J.-P. Tixier, Notes on Deep Learning for NLP, 2018