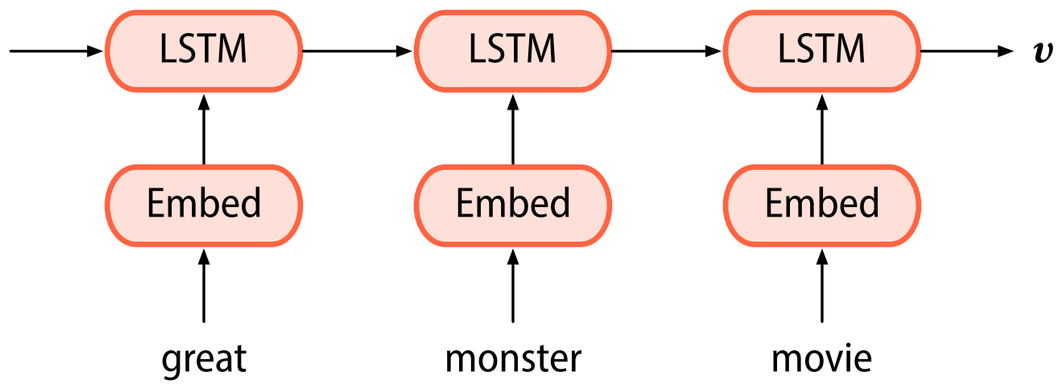
Deep Learning for Natural Language Processing

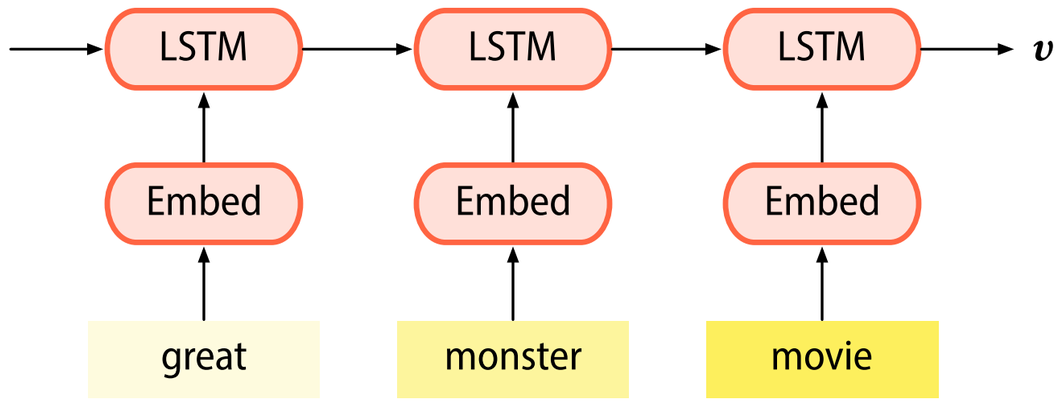
Attention

Marco Kuhlmann
Department of Computer and Information Science
Linköping University

Recency bias in recurrent neural networks



Recency bias in recurrent neural networks



The last hidden state is prone to bias towards the recent past.

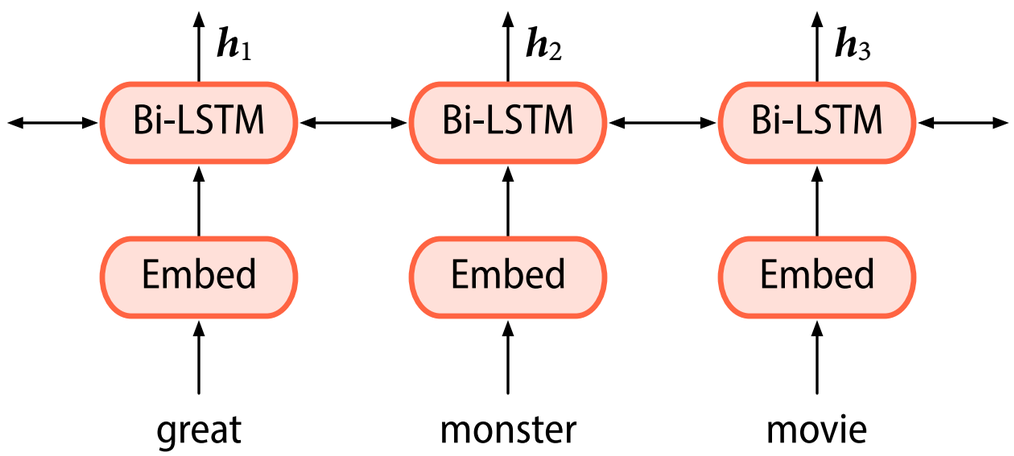
Chen et al. (2016); Werlen et al. (2018)

Attention

- In the context of text classification, **attention** enables the model to learn which words are the most important ones.
- Essentially, we compute a set of weights that allow us to score words based on how much the model should 'attend to them'.
- Attention was first proposed in the context of neural machine translation, but is now used in many models.

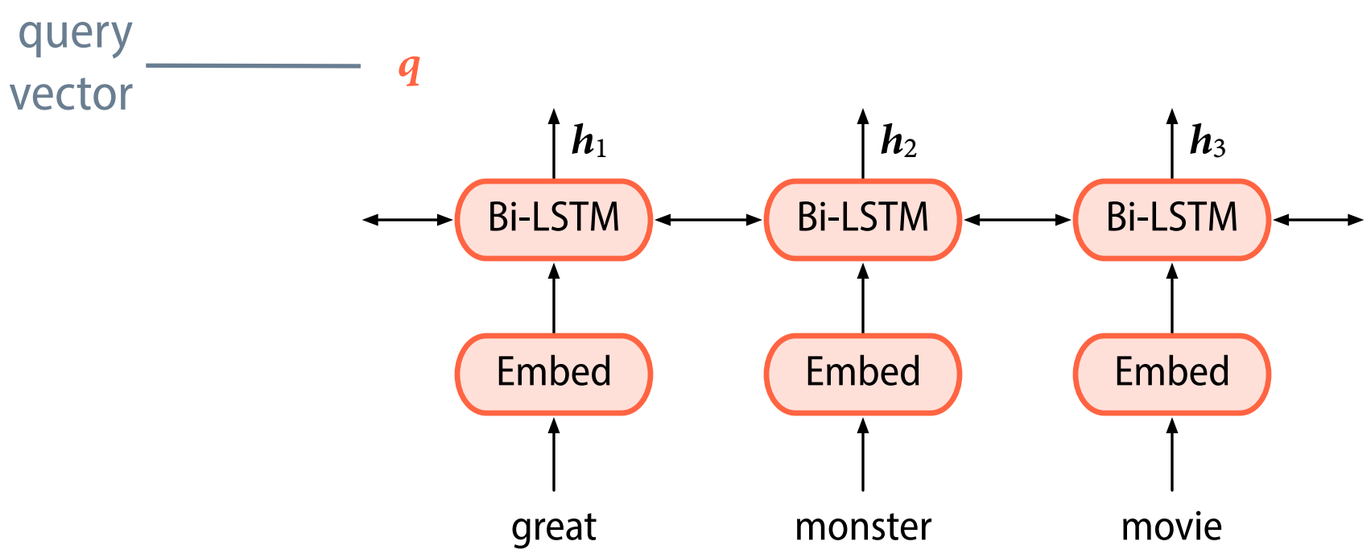
  Bahdanau et al. (2015)

Attention for classification



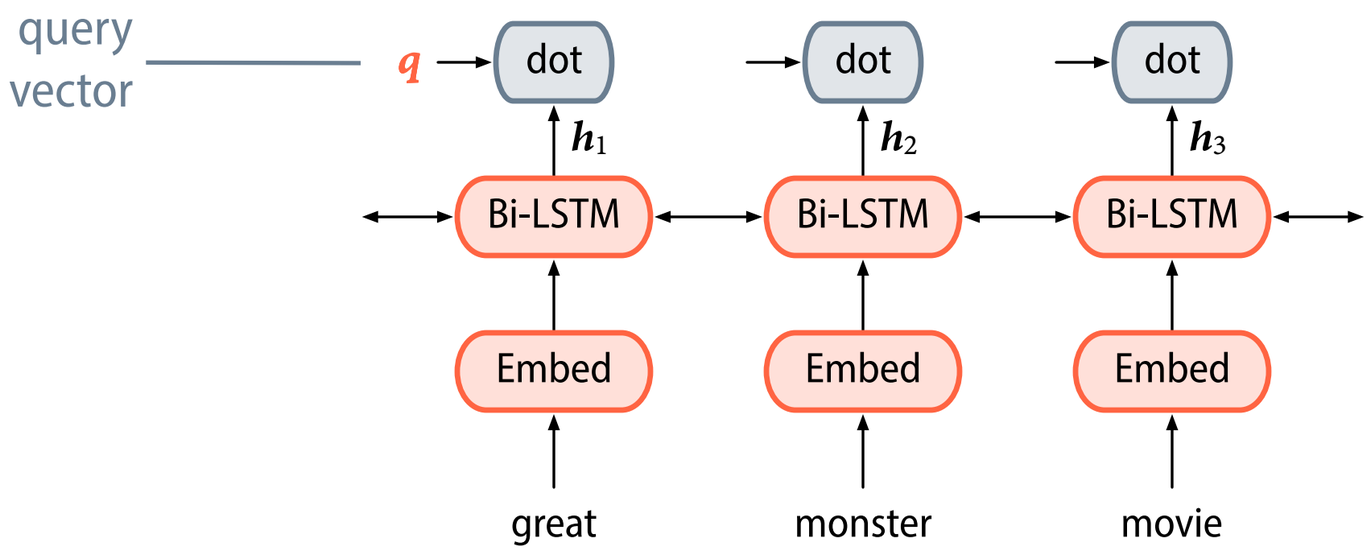
Cheng et al. (2016)

Attention for classification



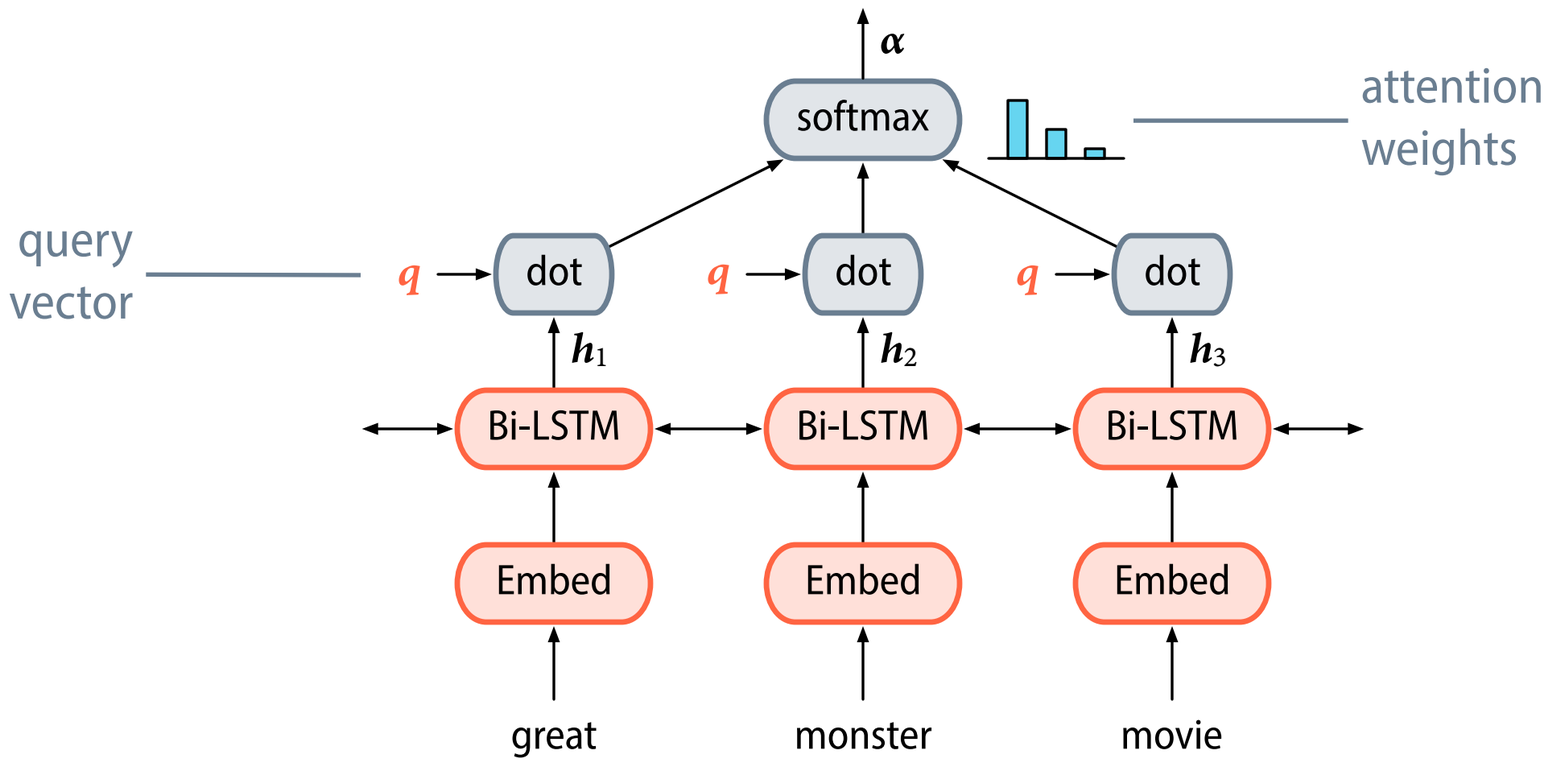
Cheng et al. (2016)

Attention for classification



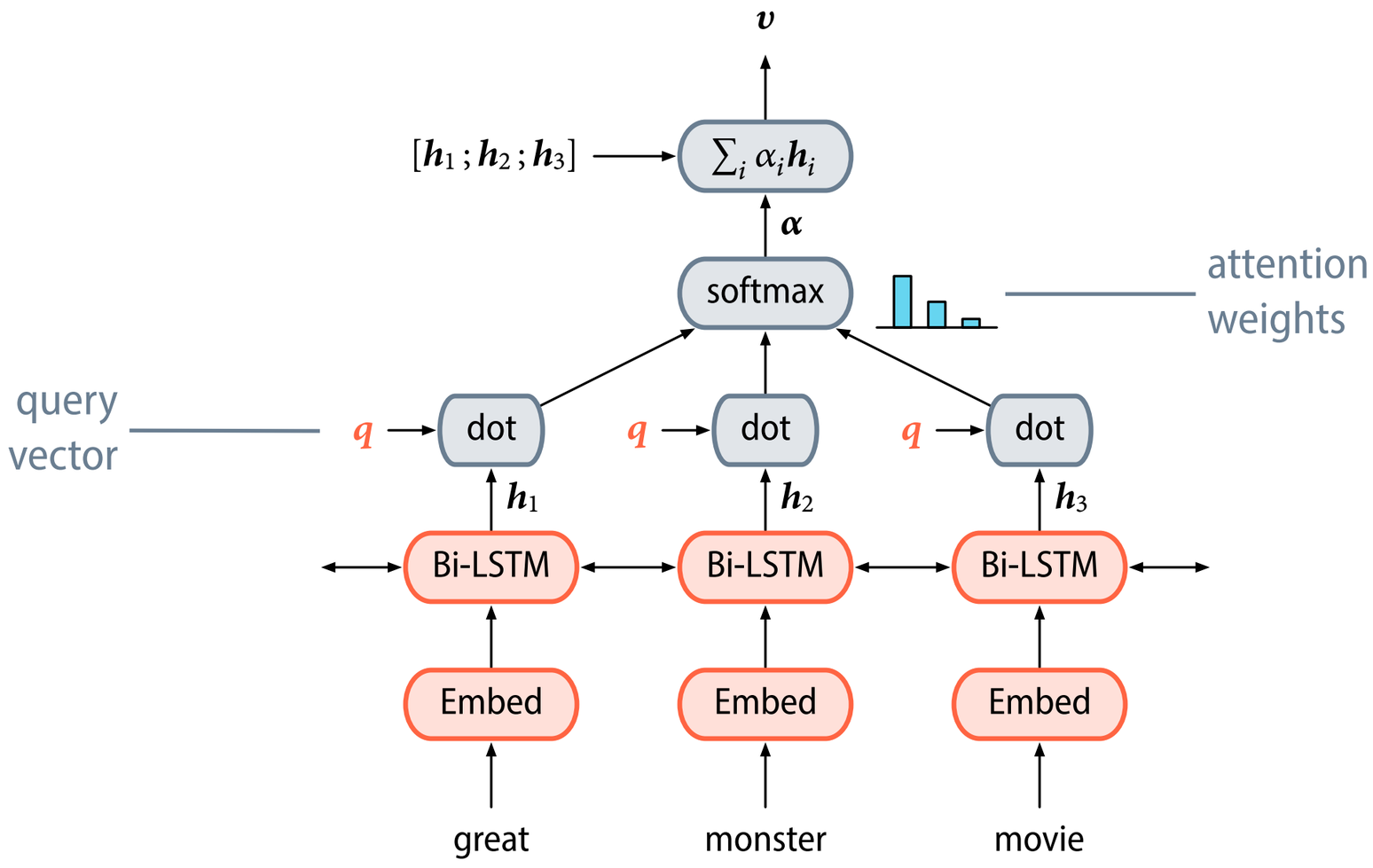
Cheng et al. (2016)

Attention for classification



Cheng et al. (2016)

Attention for classification



Cheng et al. (2016)

A more general characterisation of attention

- In general, attention can be described as a mapping from a query $\boldsymbol{q}$ and a set of key–value pairs $\boldsymbol{k}_i$, $\boldsymbol{v}_i$ to an output.
- The output is the weighted sum of the $\boldsymbol{v}_i$, where the weight of each $\boldsymbol{v}_i$ is given by the compatibility between $\boldsymbol{q}$ and $\boldsymbol{k}_i$.

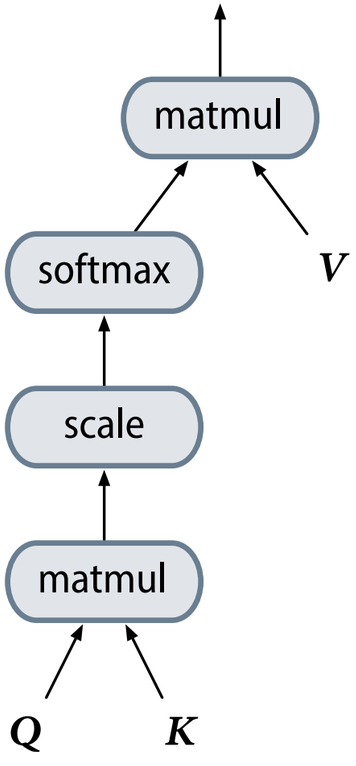
  The dot product provides a measure of compatibility.

- In the classification architecture, keys and values are the same; they correspond to the hidden states $\boldsymbol{h}_i$.

Vaswani et al. (2017)

Scaled dot-product attention

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}$$



- Used in the Transformer architecture.

  Vaswani et al. (2017)

- The input consists of queries and keys of dimension d_k, and values of dimension d_v.

- Scaling prevents the softmax from being pushed into regions with small gradients.

Interpretation of attention

- In addition to improved performance, attention is attractive because it allows us to inspect what a network attends to.

  visualise weights; correlate weights to external data or human rationales

- The discussion of the possibilities and limitations of using attention to interpret neural models is ongoing.

  Is Attention Interpretable? (Serrano and Smith, 2019)

  Attention is not Explanation (Jain and Wallace, 2019)

  Attention is not not Explanation (Wiegreffe and Pinter, 2019)