Deep Learning for Natural Language Processing

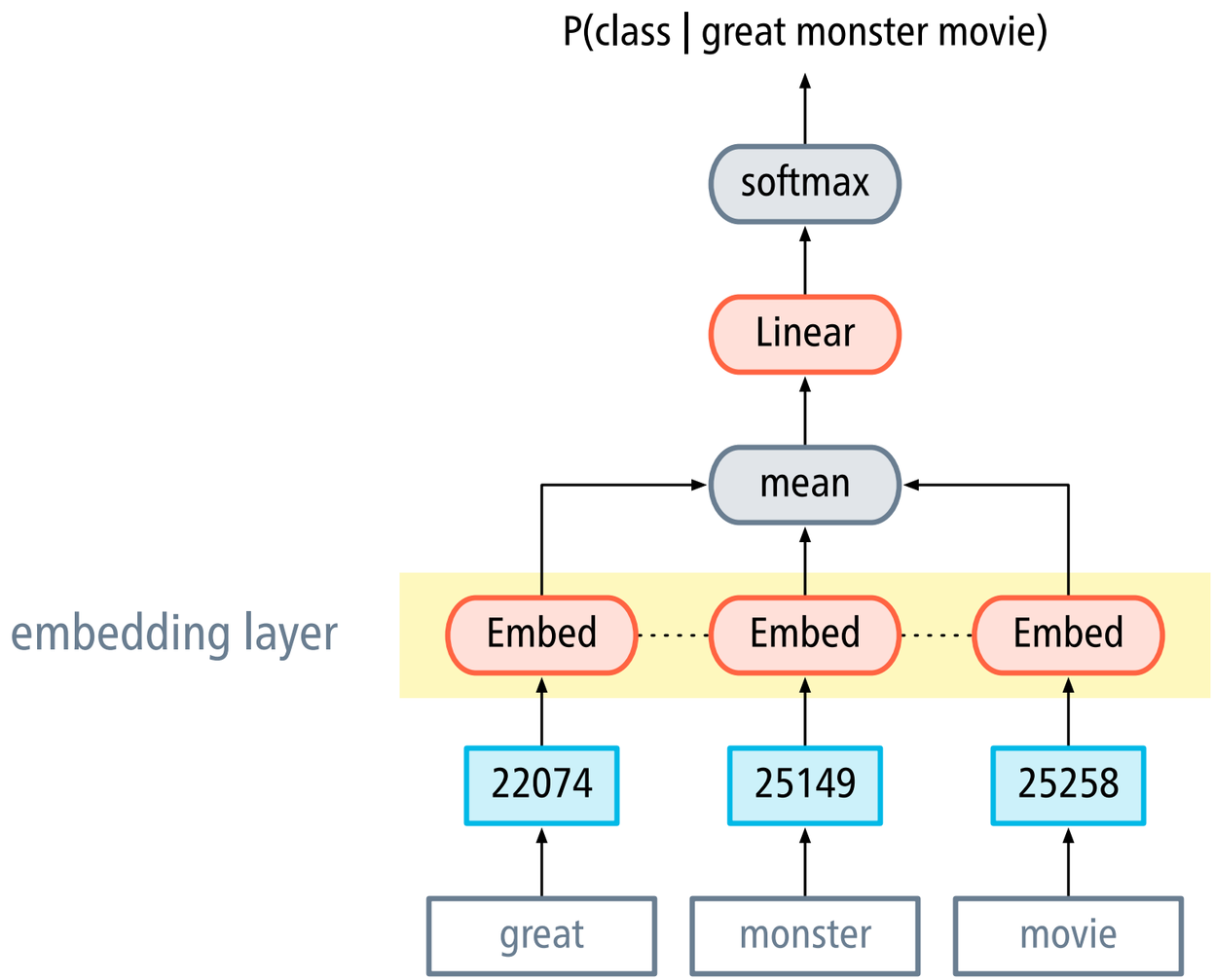
Recap Unit 1 & Unit 2

Marco Kuhlmann
Department of Computer and Information Science

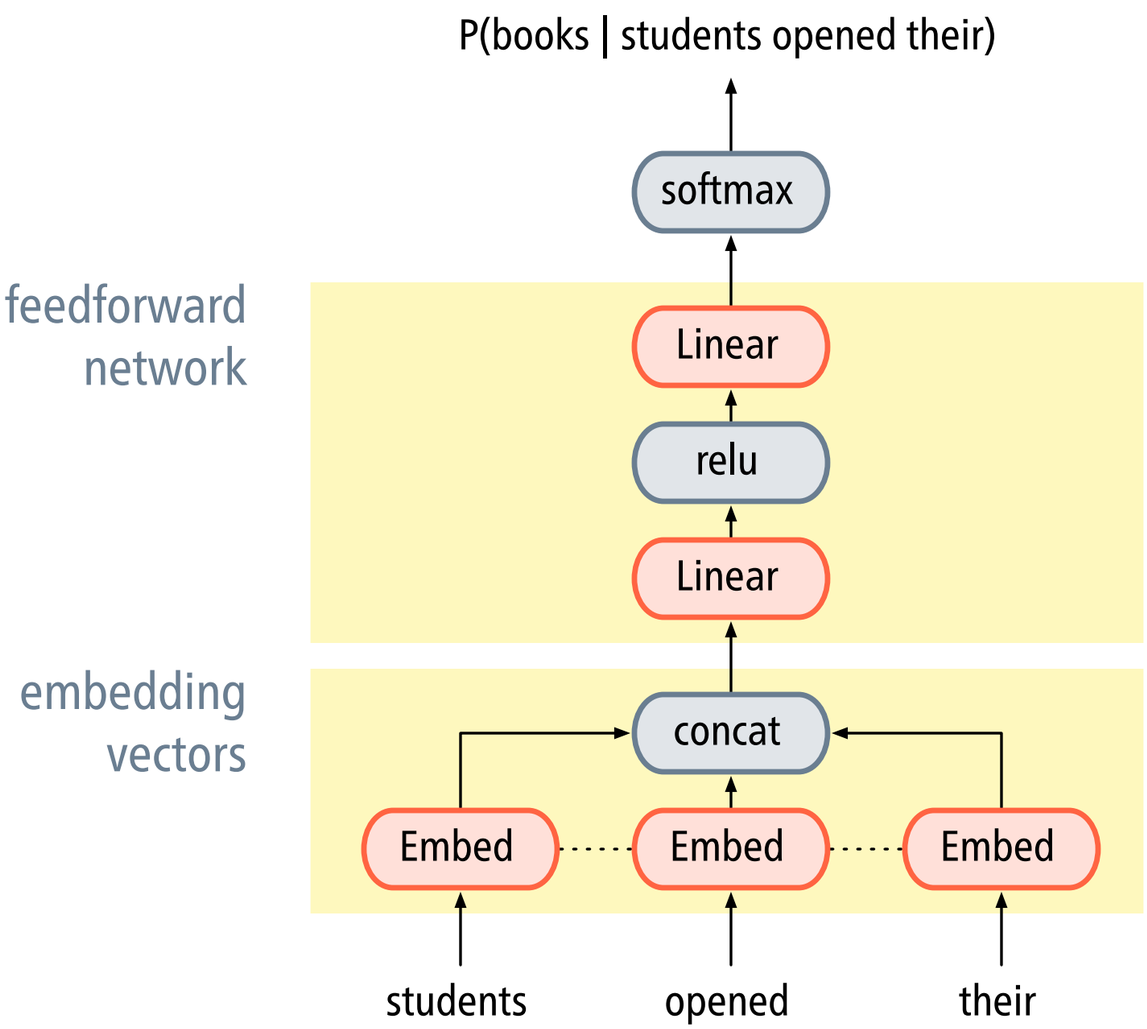
LINKÖPING UNIVERSITY

Warming up: Counting the number of parameters

Bag-of-words classifier



A neural four-gram model

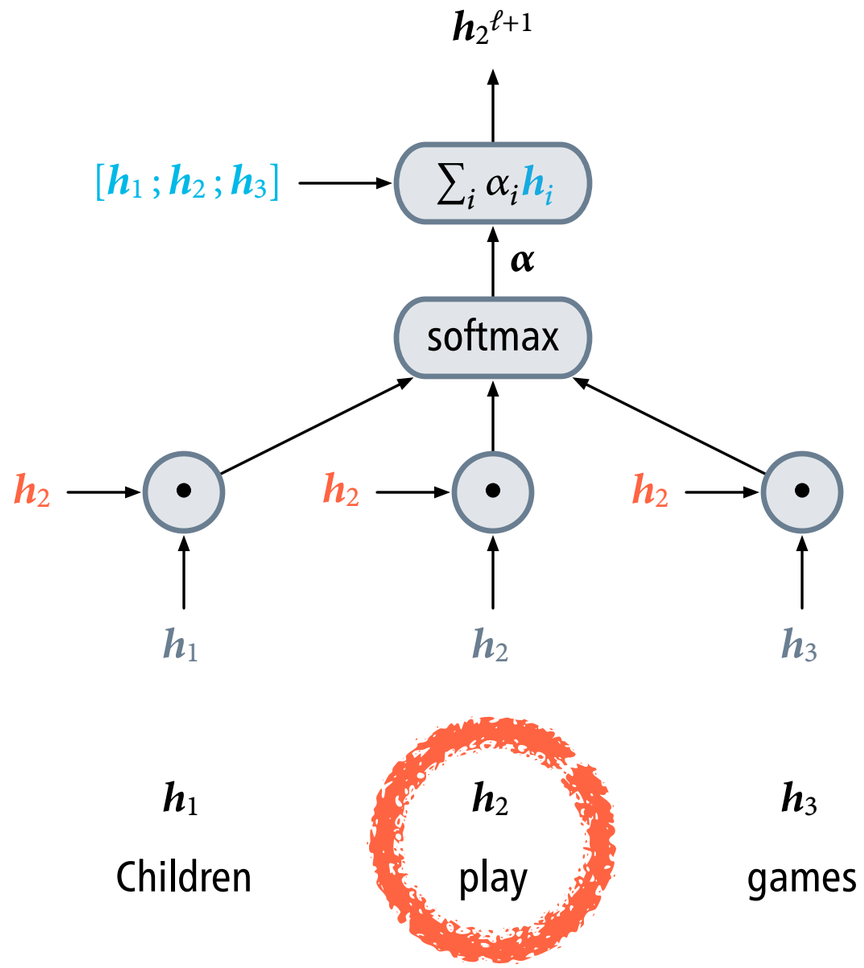


Overview of Unit 1

Overview of Unit 2

Muddiest point

Contextual embeddings via attention



Attention

Consider the following values for the attention example:

$\boldsymbol{h}_1 = [0.5539, 0.7239]$ $\boldsymbol{h}_2 = [0.4111, 0.3878]$ $\boldsymbol{h}_3 = [0.2376, 0.1264]$

Assuming that the attention score is computed using the (unscaled) dot product, what is the refined representation for $\boldsymbol{h}_2$?

A. [0.3962, 0.3279, 0.2759]

B. [0.4198, 0.4488]

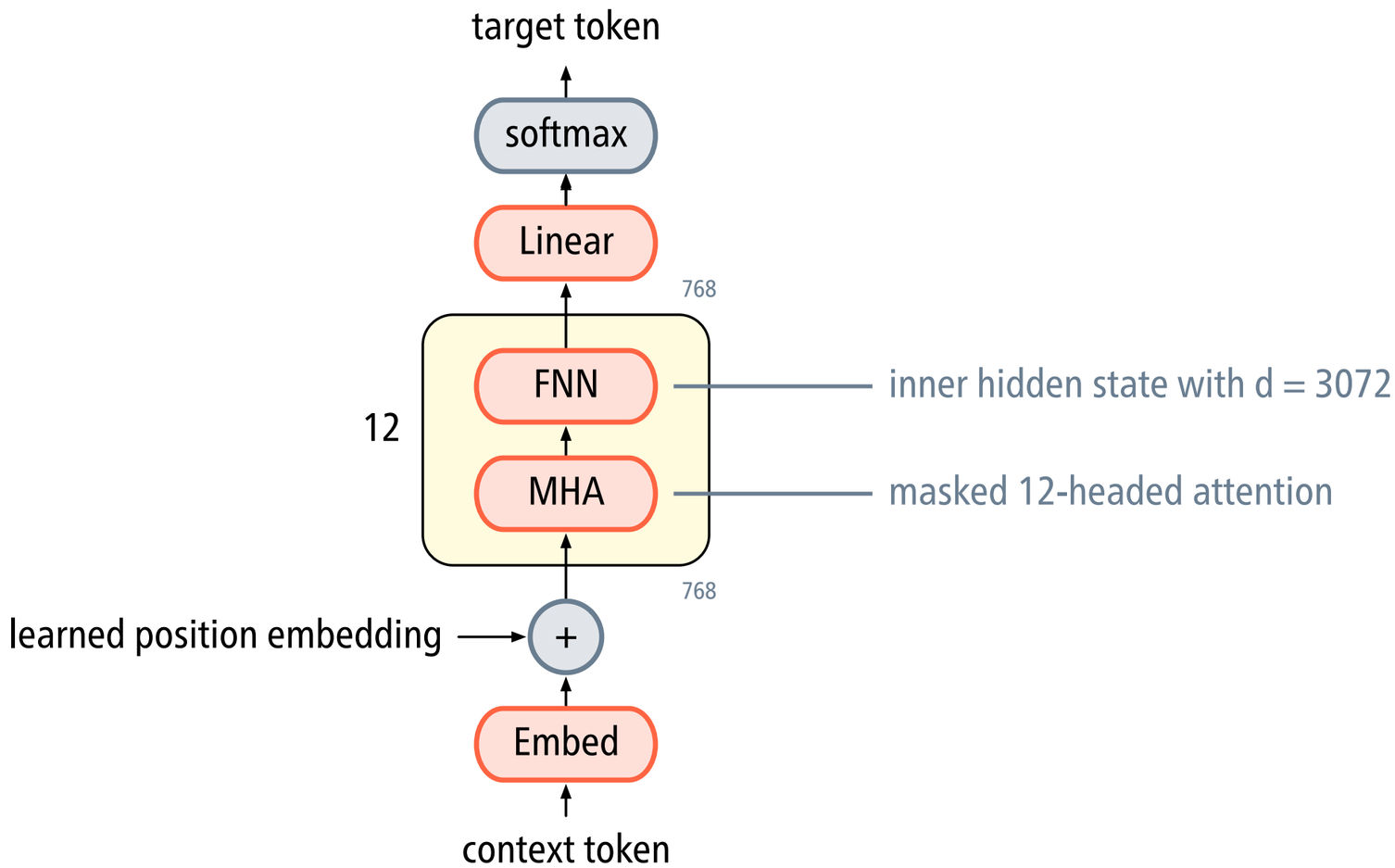
C. [0.5084, 0.3194, 0.1467]

Attention

Which of the following statements about the characterisation of attention in terms of queries, keys, and values is true?

- The query has the same length as each value.
- The output has the same length as each value.
- Each key has the same length as each value.

GPT model architecture



Radford et al. (2018)