

Deep Learning for Natural Language Processing

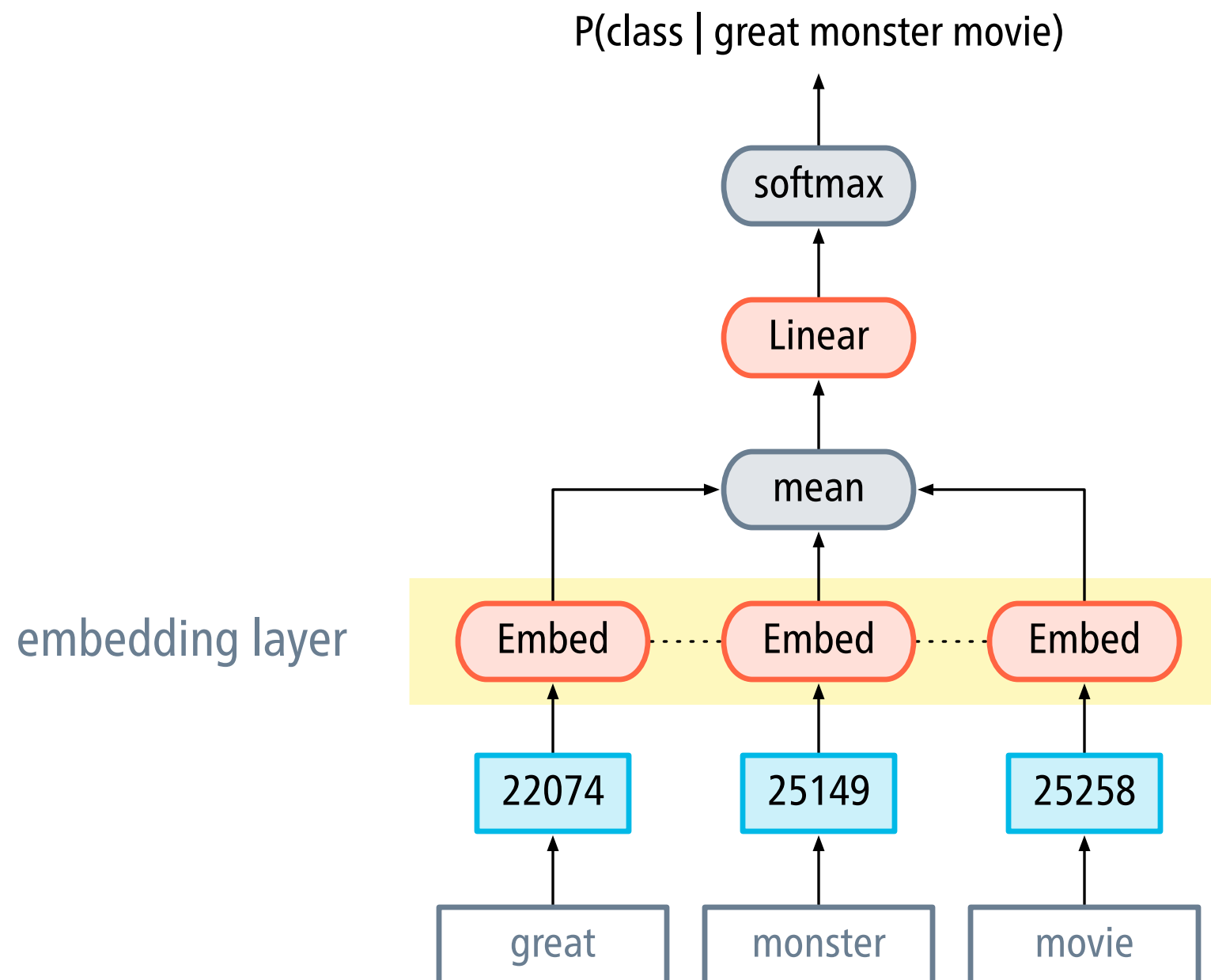
Recap Unit 1 & Unit 2

Marco Kuhlmann

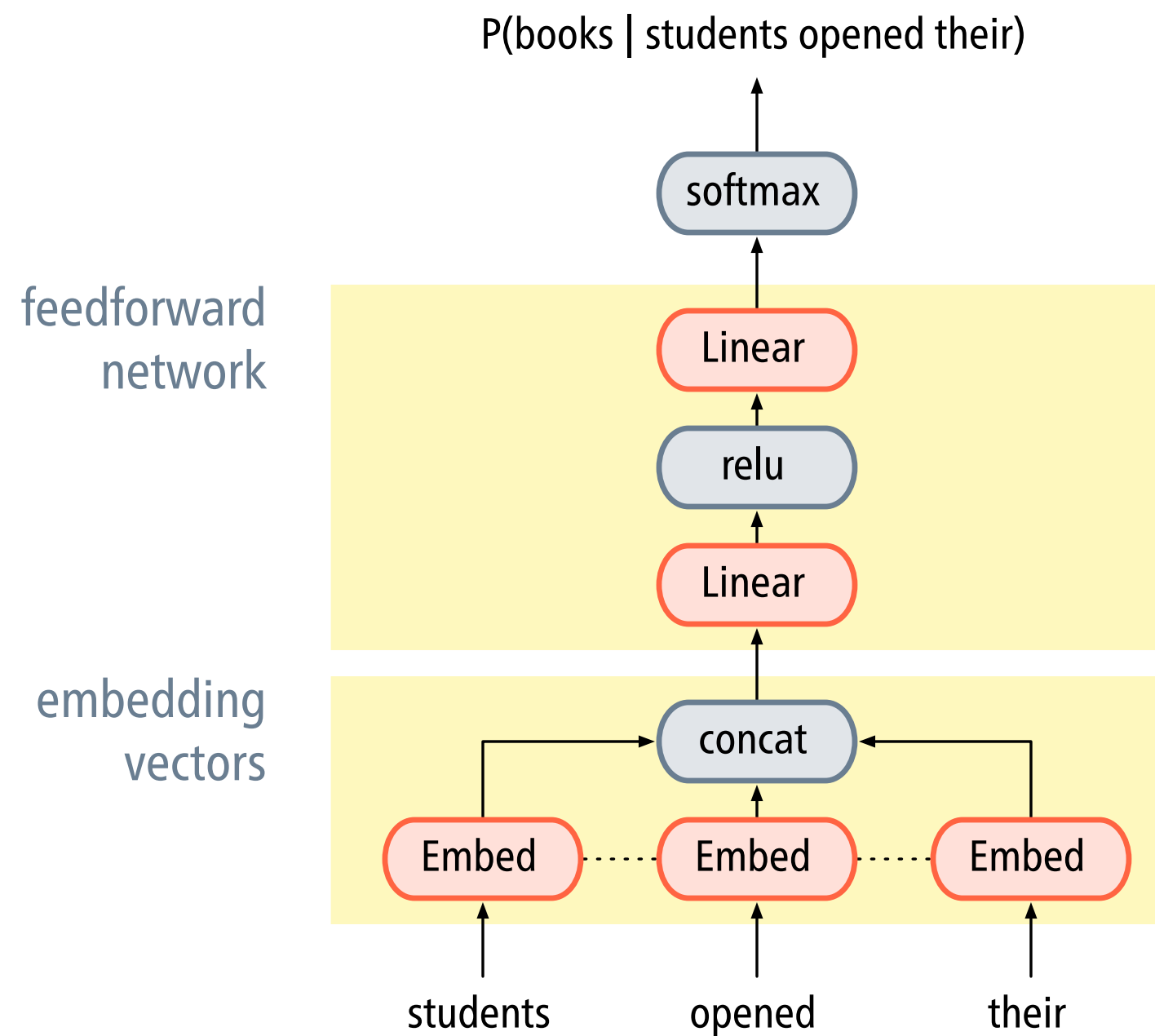
Department of Computer and Information Science

Warming up: Counting the number of parameters

Bag-of-words classifier



A neural four-gram model



Overview of Unit 1

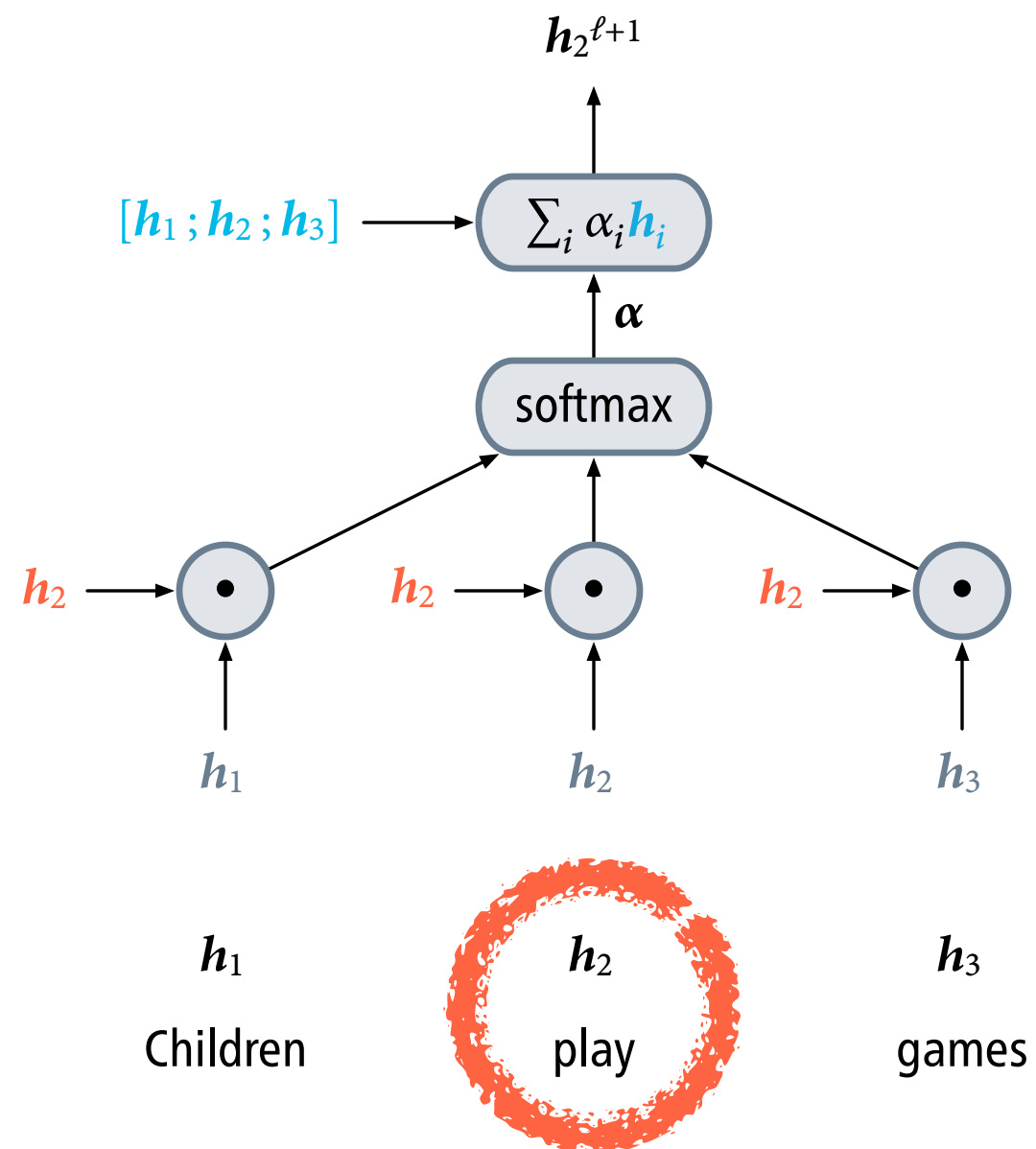
- 1.1 Introduction to tokenisation
- 1.2 The Byte Pair Encoding algorithm
- 1.3 Tokenisation fairness
- 1.4 Introduction to embeddings
- 1.5 Word embeddings
- 1.6 Contextualised word embeddings

Overview of Unit 2

- 2.1 Attention
- 2.2 Introduction to Transformers
- 2.3 Transformers in more detail
- 2.4 Representing positions in Transformers
- 2.5 Generating text from a language model
- 2.6 Transformer representation models

Muddiest point

Contextual embeddings via attention



Attention

Consider the following values for the attention example:

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

Assuming that the attention score is computed using the (unscaled) dot product, what is the refined representation for \mathbf{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

