BERT:

Pre-training of Deep Bidirectional Transformers for Language Understanding

Vasiliki Kougia
Instructor: John Pavlopoulos
28/4/2021

Modern NLP reading group

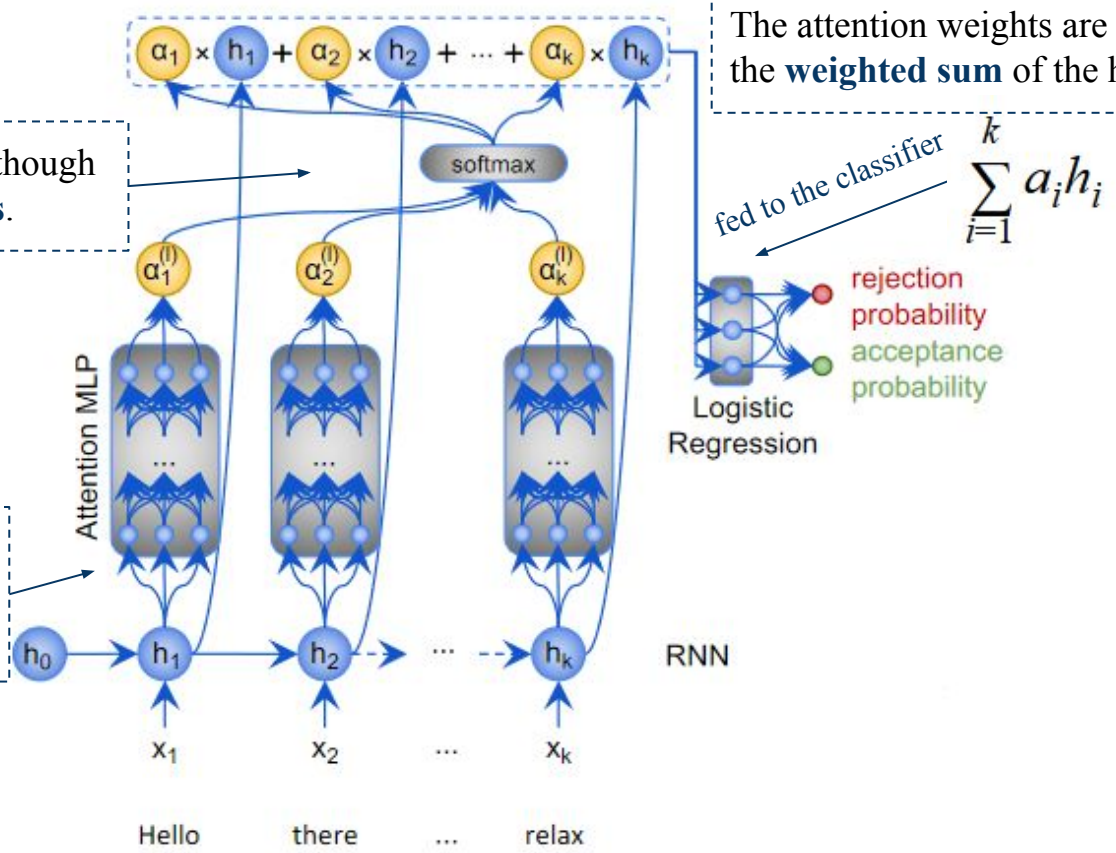
Stockholm University

Self-attention in RNNs

The attention weights are used to obtain the **weighted sum** of the hidden states.

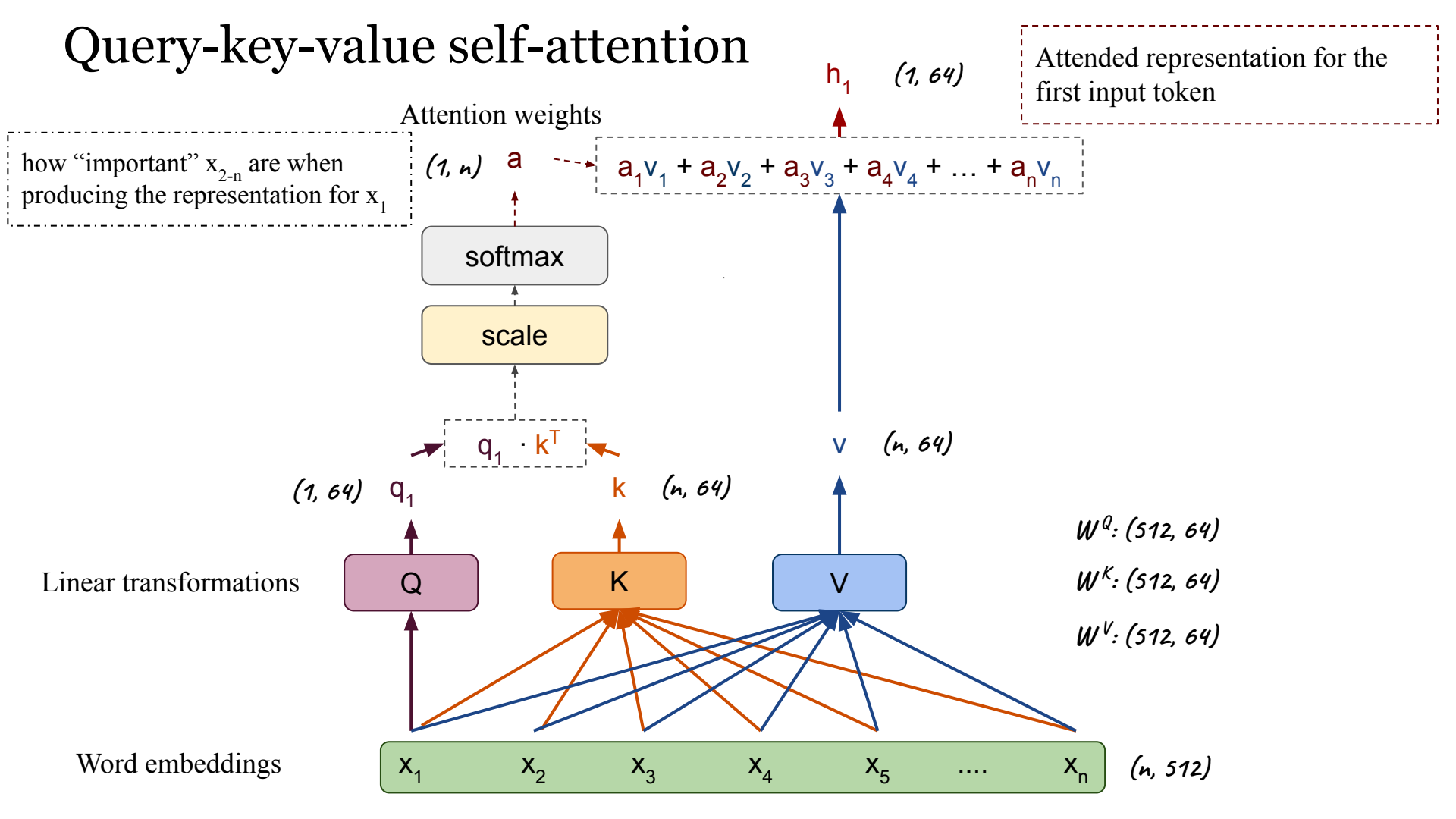
The outputs of the MLP are passed though softmax to obtain **attention weights**.

The hidden states of an RNN are passed through the **attention MLP**.

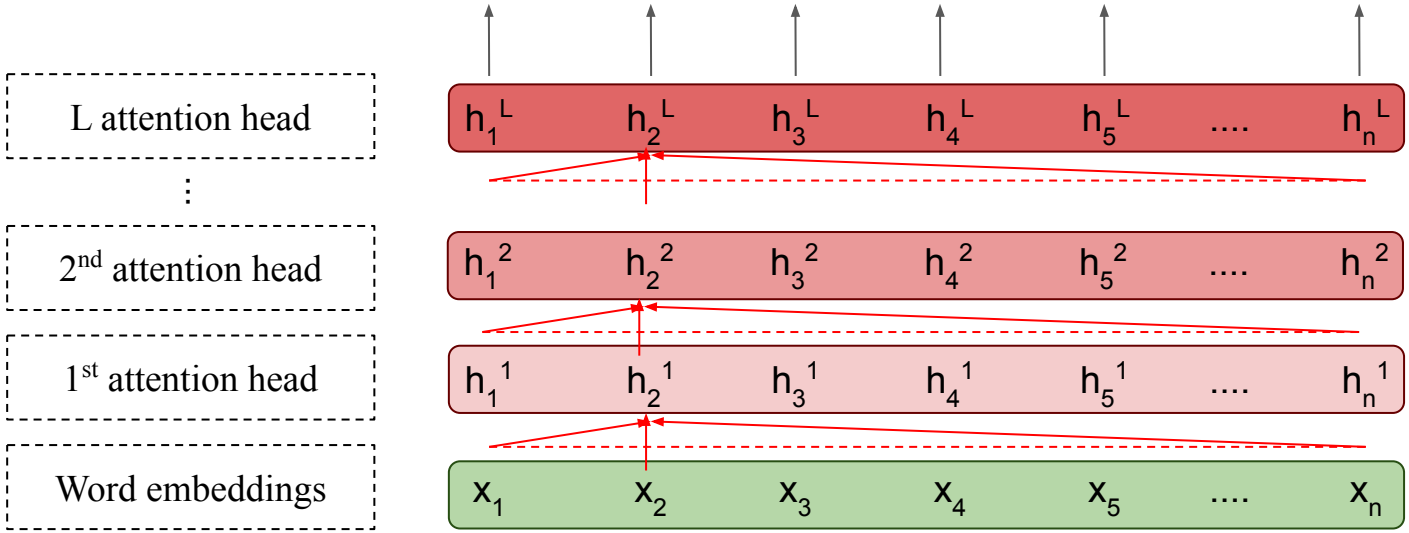


J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deeper Attention to Abusive User Content Moderation", EMNLP 2017, http://nlp.cs.aueb.gr/pubs/emnlp2017.pdf

Query-key-value self-attention



Transformers - the bigger picture



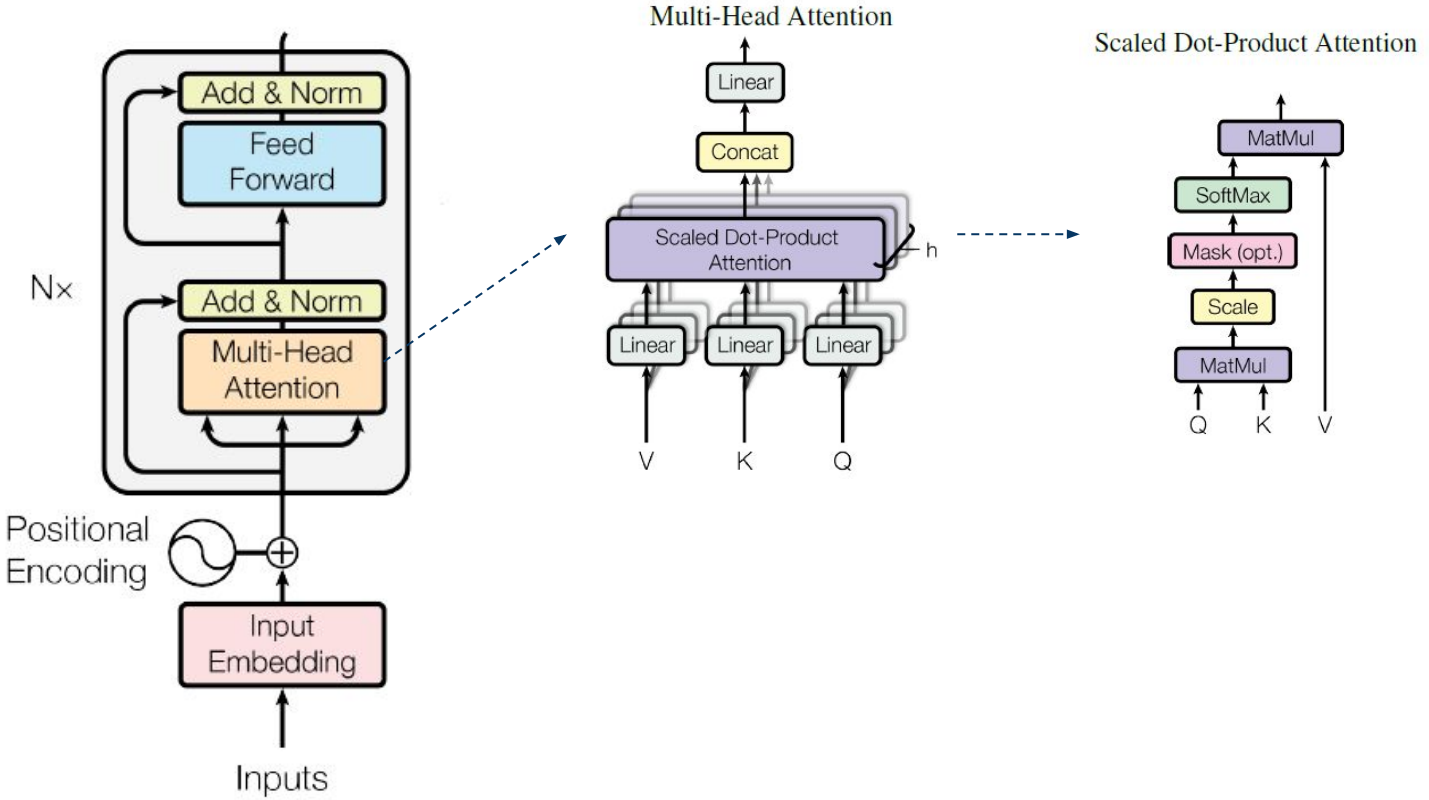
For position 2

$$h_i^j = \sum_{r=1}^{n} softmax(q_i^j \, k_r^{j\,T}) \, v_r^j$$

$$q_i^j = W^{Q,j} \, h_i^{j-1}$$
$$k_r^j = W^{K,j} \, h_r^{j-1}$$
$$v_r^j = W^{V,j} \, h_r^{j-1}$$

Different weight matrices at each head

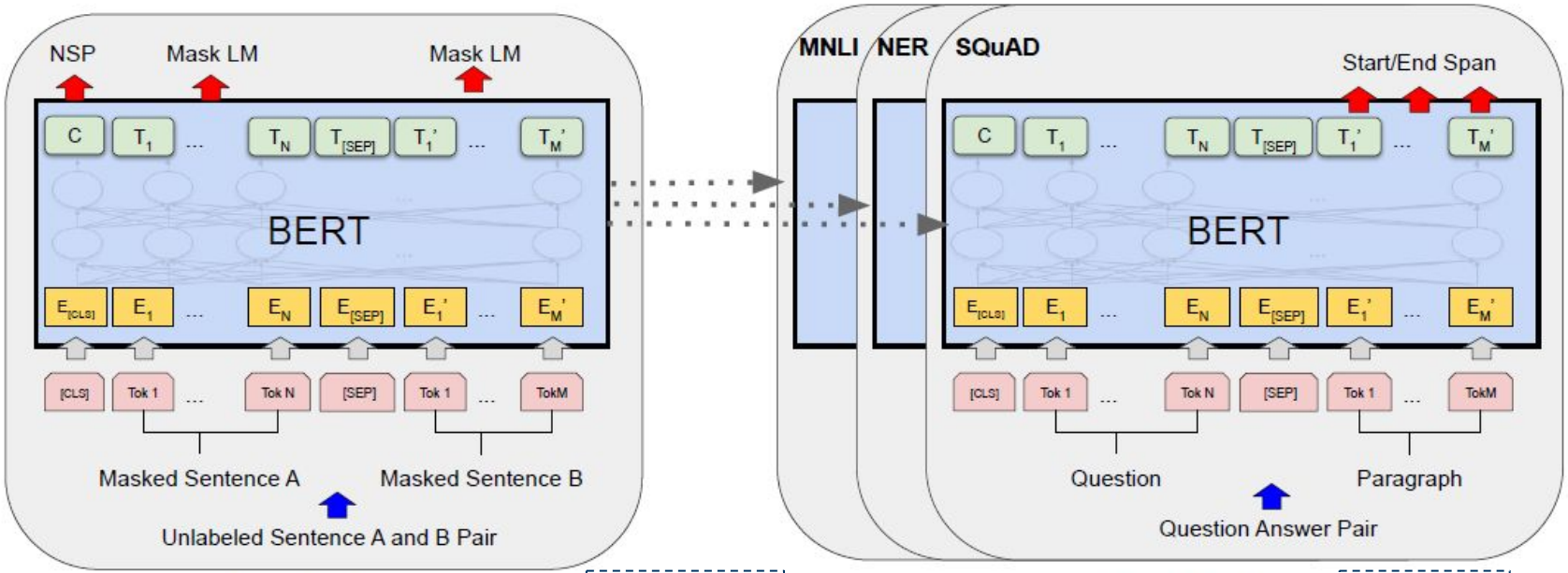
Transformer encoder block



Vaswani et al., “Attention is All You Need”, 2017, https://arxiv.org/abs/1706.03762

BERT

- Consists of stacked transformer encoders
- Pre-trained on a huge corpus



Pre-training

unsupervised

Fine-Tuning

supervised

For Masked Language Modeling (MLM) and Next Sentence Prediction (NSP)

For different down-stream tasks

Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018, https://arxiv.org/abs/1810.04805

BERT - text classification