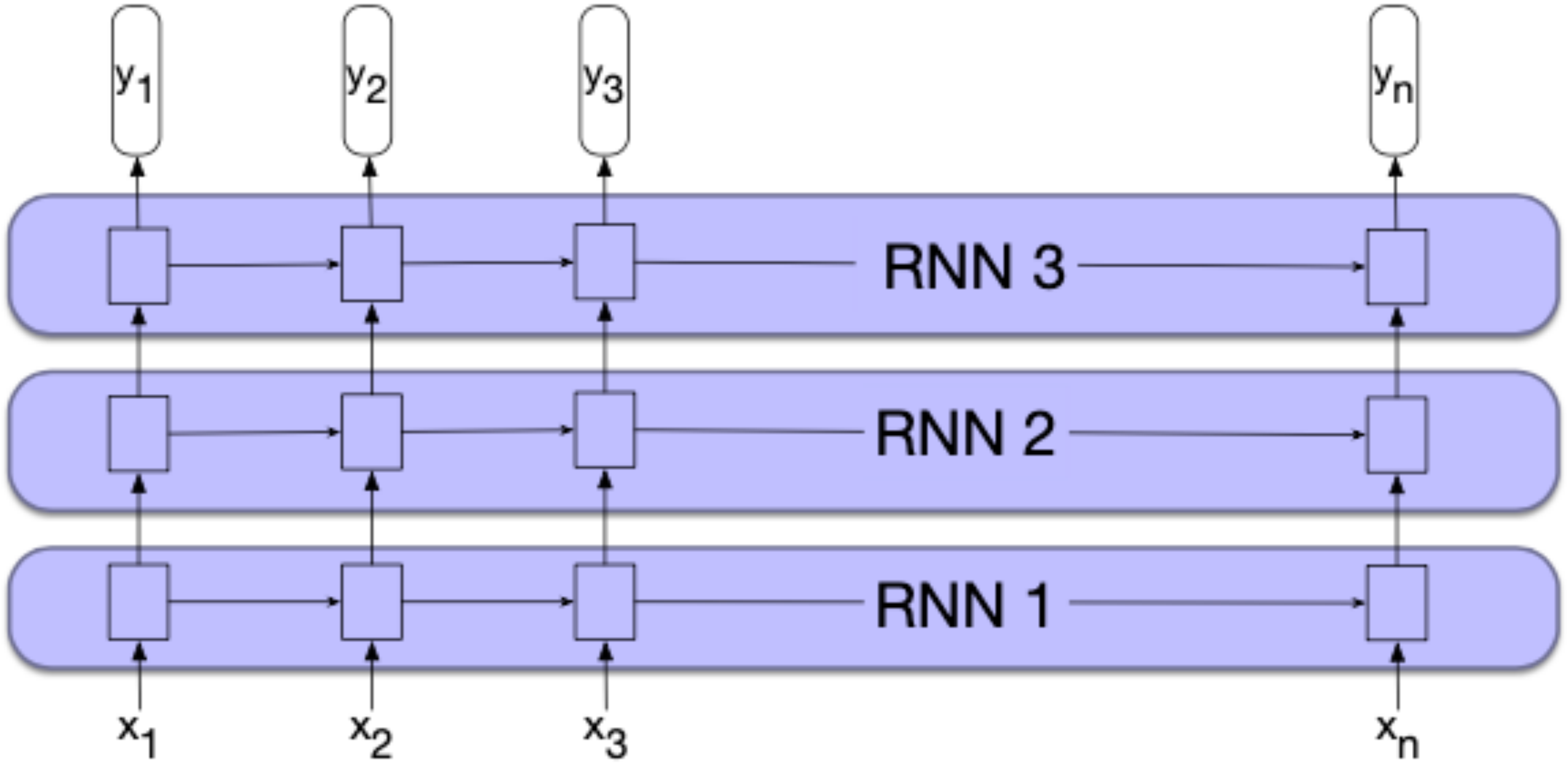
22-10–2024

RNN and Self-Attention



Stacked RNN

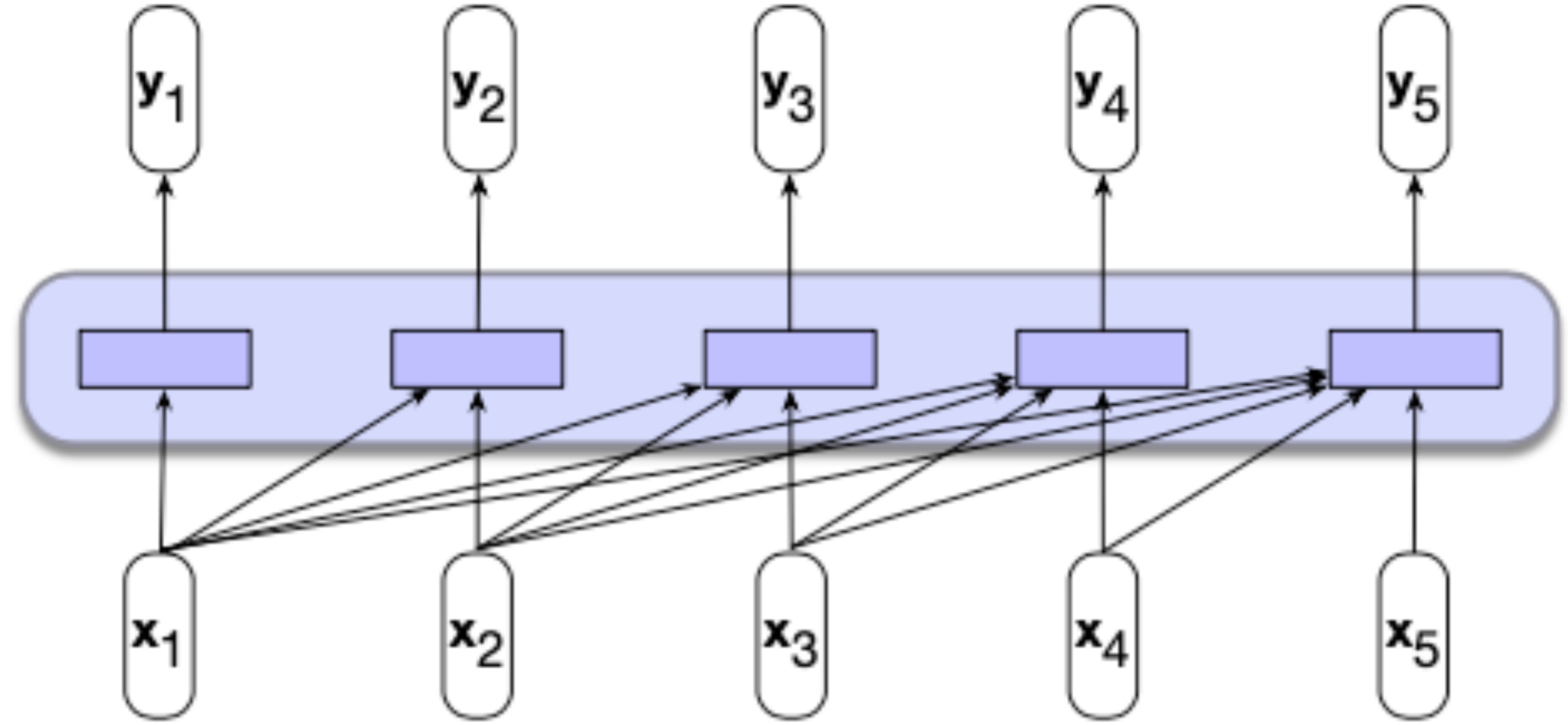


Image: Jurafsky & Martin “Speech and Language Processing, 3rd ed., 2023

Self-Attention

- Operations in self attention:
  - $y_3 = [x_3 \cdot x_3]x_3 + [x_3 \cdot x_2]x_2 + [x_3 \cdot x_1]x_1$
  - $y_i = \sum_{j=1}^{i} [x_i \cdot x_j]x_j$
  - $y_i = \sum_{j=1}^{i} [(W^Q x_i) \cdot (W^K x_j)](W^V x_j)$
  - $X = \left[x_1; x_2; \cdots; x_N\right] \in \mathbf{R}^{N \times d}$, $W^Q \in \mathbf{R}^{d \times d}$,
  - $W^K \in \mathbf{R}^{d \times d}$, $W^V \in \mathbf{R}^{d \times d}$
  - $Y = \left[(XW^Q)(XW^K)^T\right](XW^V) \in \mathbf{R}^{N \times d}$
  - $Y = softmax\left(\frac{QK^T}{d}\right)V$

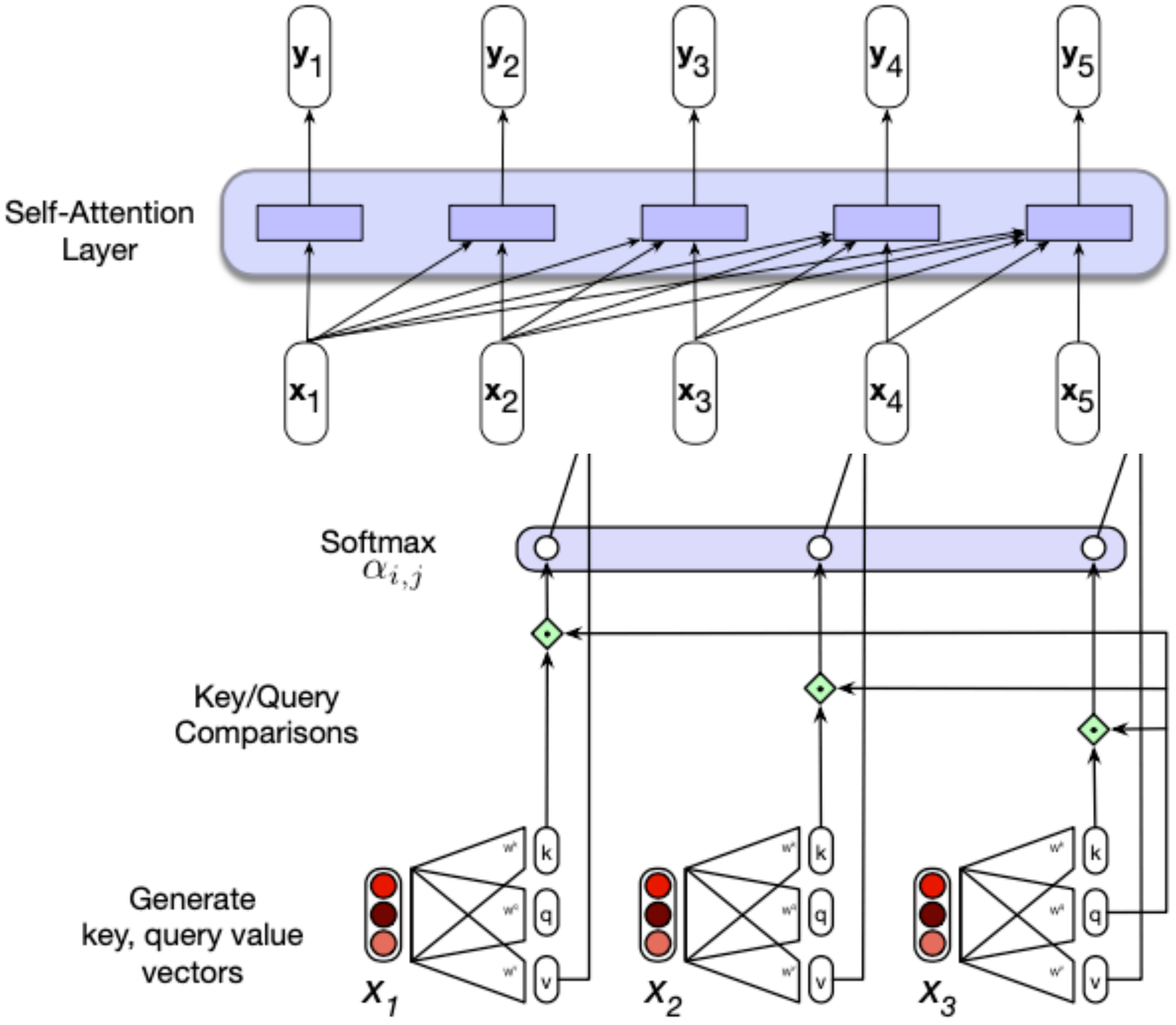
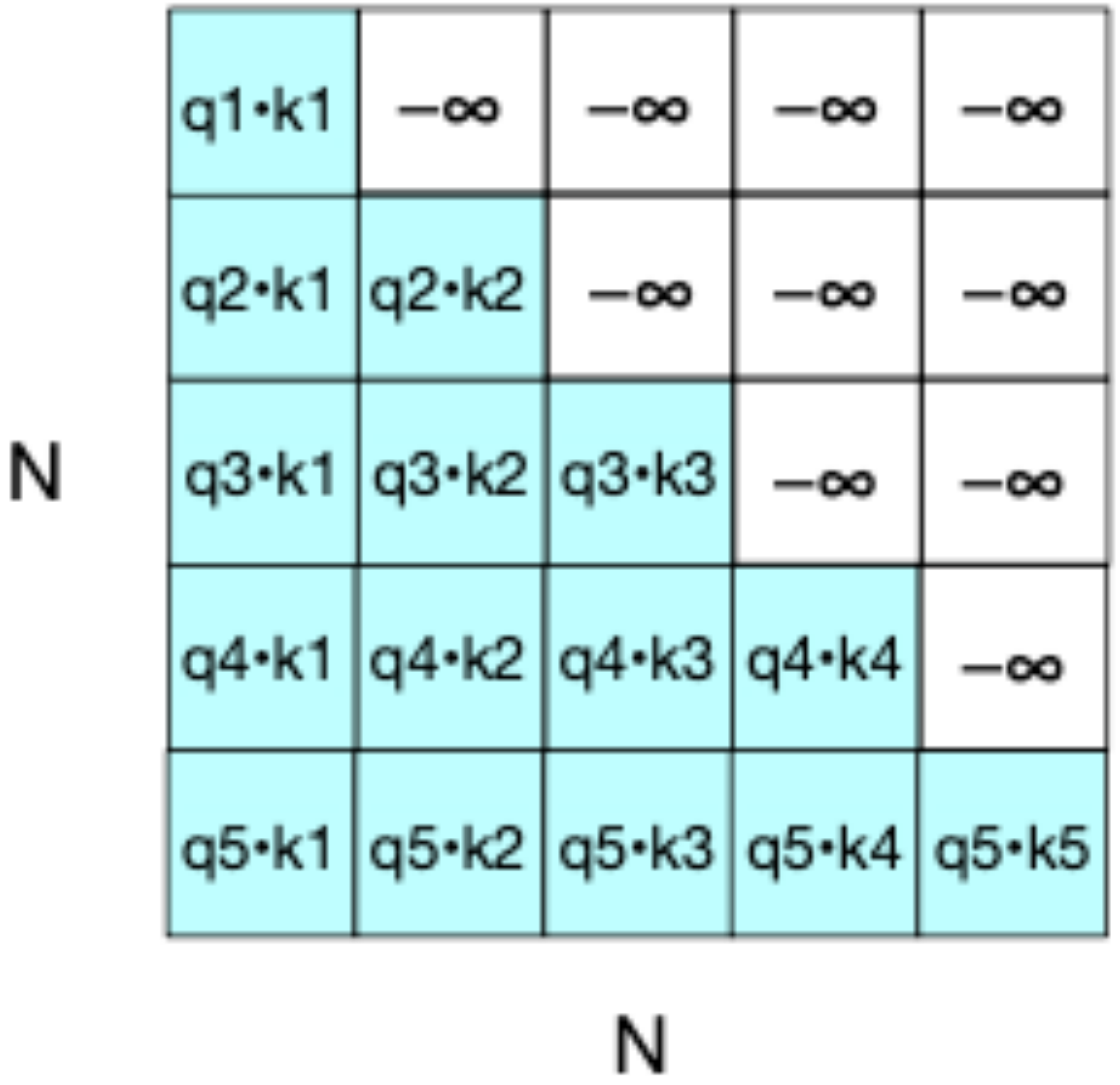


Image: Jurafsky & Martin “Speech and Language Processing, 3rd ed., 2023

Self-Attention (cont.)

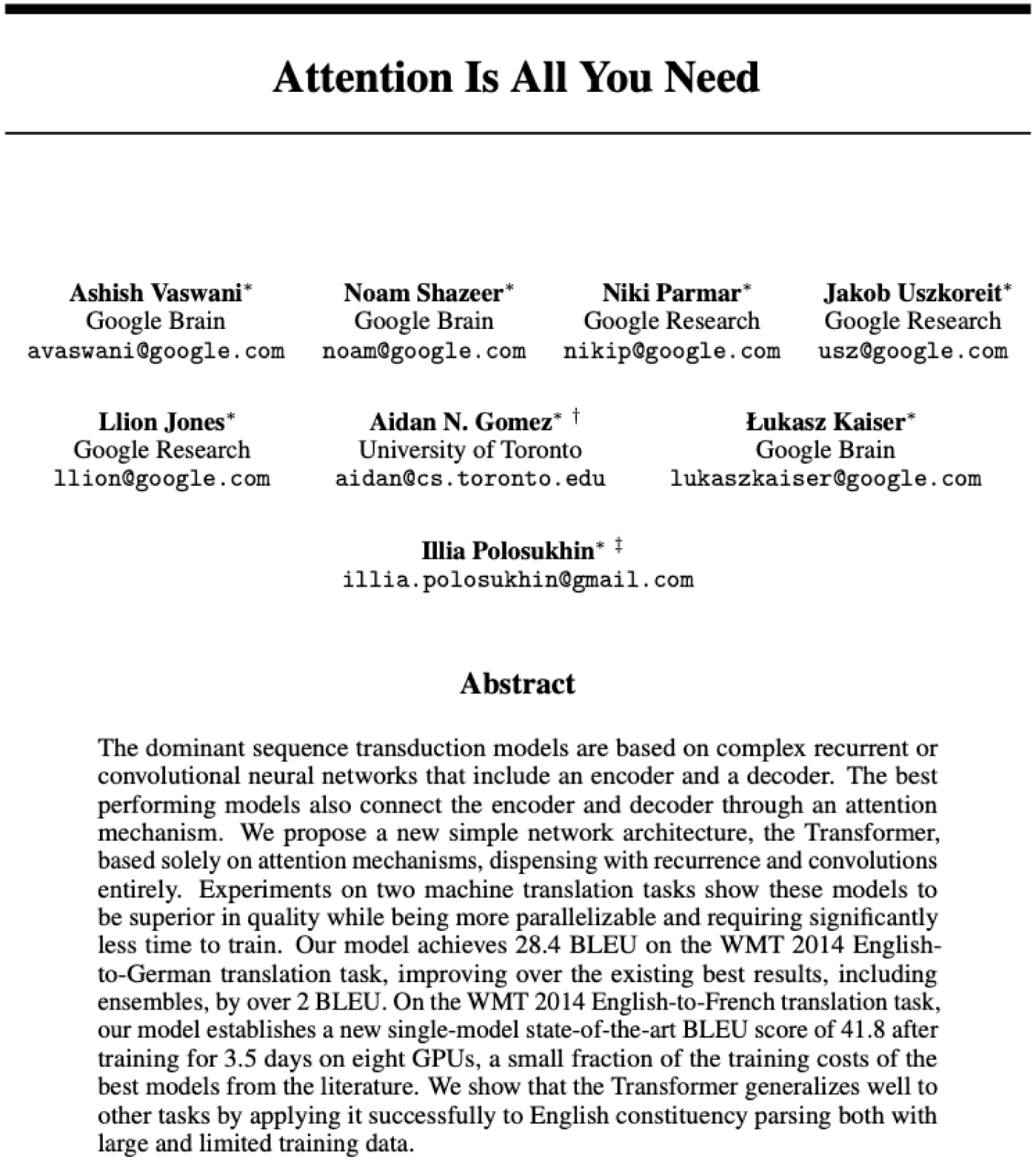
- $X = \left[x_1; x_2; \cdots; x_N\right] \in \mathbf{R}^{N \times d}$, $W^Q \in \mathbf{R}^{d \times d}$,
- $W^K \in \mathbf{R}^{d \times d}$, $W^V \in \mathbf{R}^{d \times d}$
- $Y = \left[(XW^Q)(XW^K)^T\right](XW^V) \in \mathbf{R}^{N \times d}$
- $Y = softmax\left(\frac{QK^T}{d}\right)V$



q1•k1	−∞	−∞	−∞	−∞
q2•k1	q2•k2	−∞	−∞	−∞
q3•k1	q3•k2	q3•k3	−∞	−∞
q4•k1	q4•k2	q4•k3	q4•k4	−∞
q5•k1	q5•k2	q5•k3	q5•k4	q5•k5

Image: Jurafsky & Martin “Speech and Language Processing, 3rd ed., 2023

Self-Attention network: transformer



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

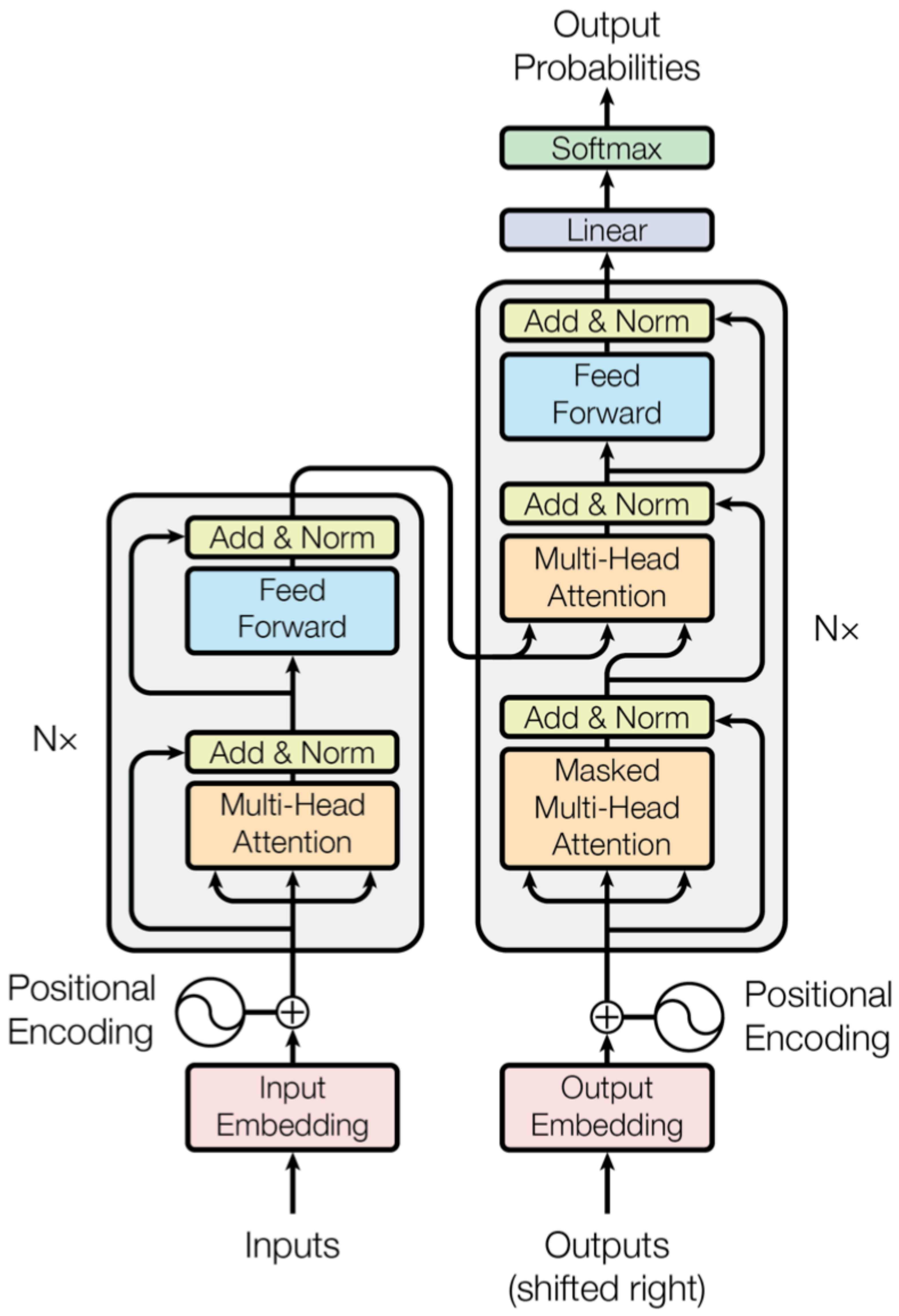
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.



Transformer block

- Layer normalisation (LN):
  - $LN_1[x + self - attention(x)]$
  - $LN_2[x + FFNN(x)]$
  - $FFNN(x) = max[0, XW_1 + b_1]W_2 + b_2$
  - $LN(\hat{x}) = \gamma\hat{x} + \beta$
    - $\hat{x} = \frac{x - \mu}{\sigma}, \mu = \frac{1}{d}\sum_{i=1}^{d} x_i; x \in \mathbf{R}^d$
    - $\sigma = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_i - \mu)^2}$

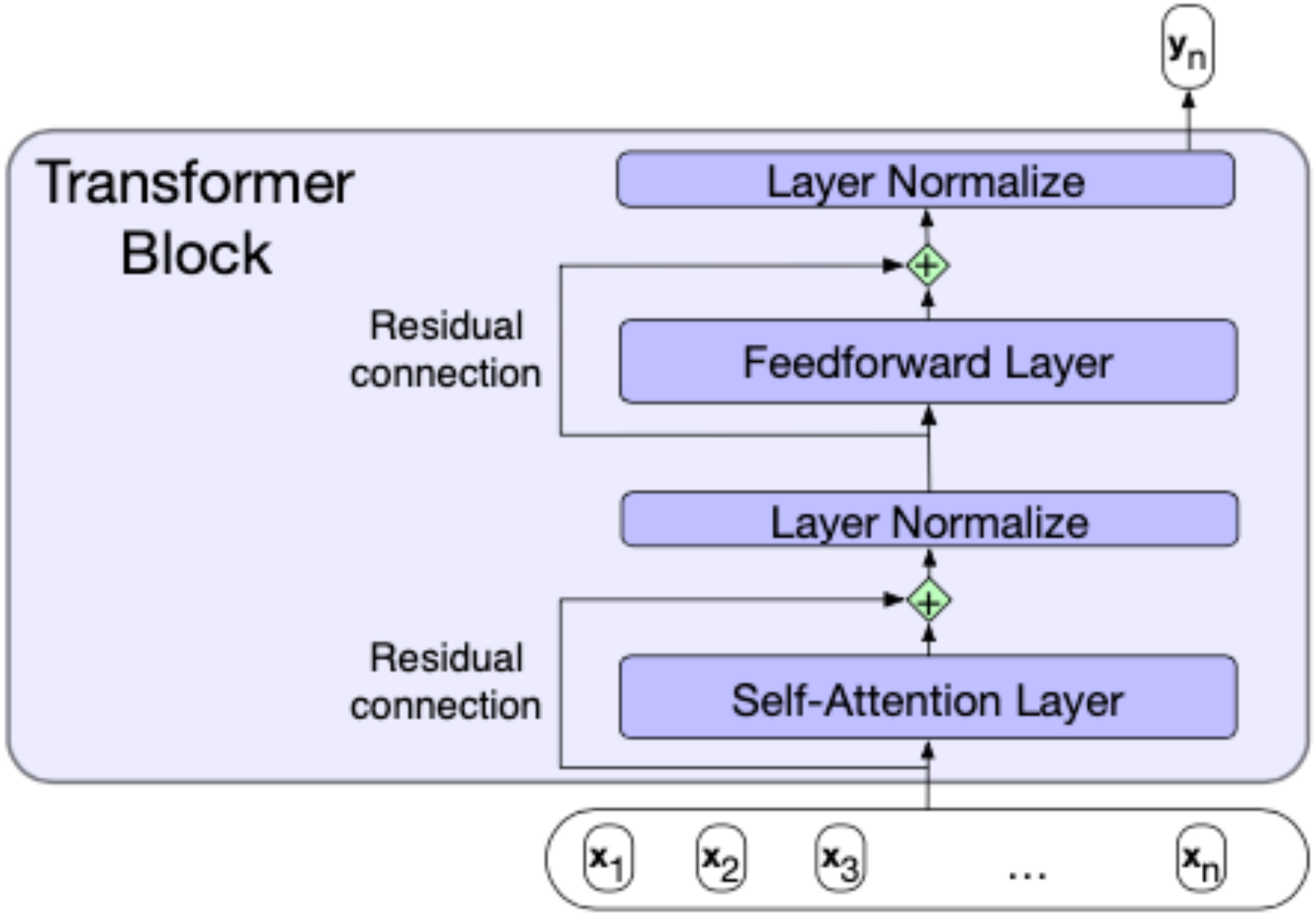


Image: Jurafsky & Martin “Speech and Language Processing, 3rd ed., 2023