Applied Deep Learning

Dr. Philippe Blaettchen
Bayes Business School (formerly Cass)

www.bayes.city.ac.uk

Learning objectives of today

Goals: Understand the basic design and functioning of transformers, a breakthrough architecture that has revolutionized NLP applications in recent years

How will we do this?

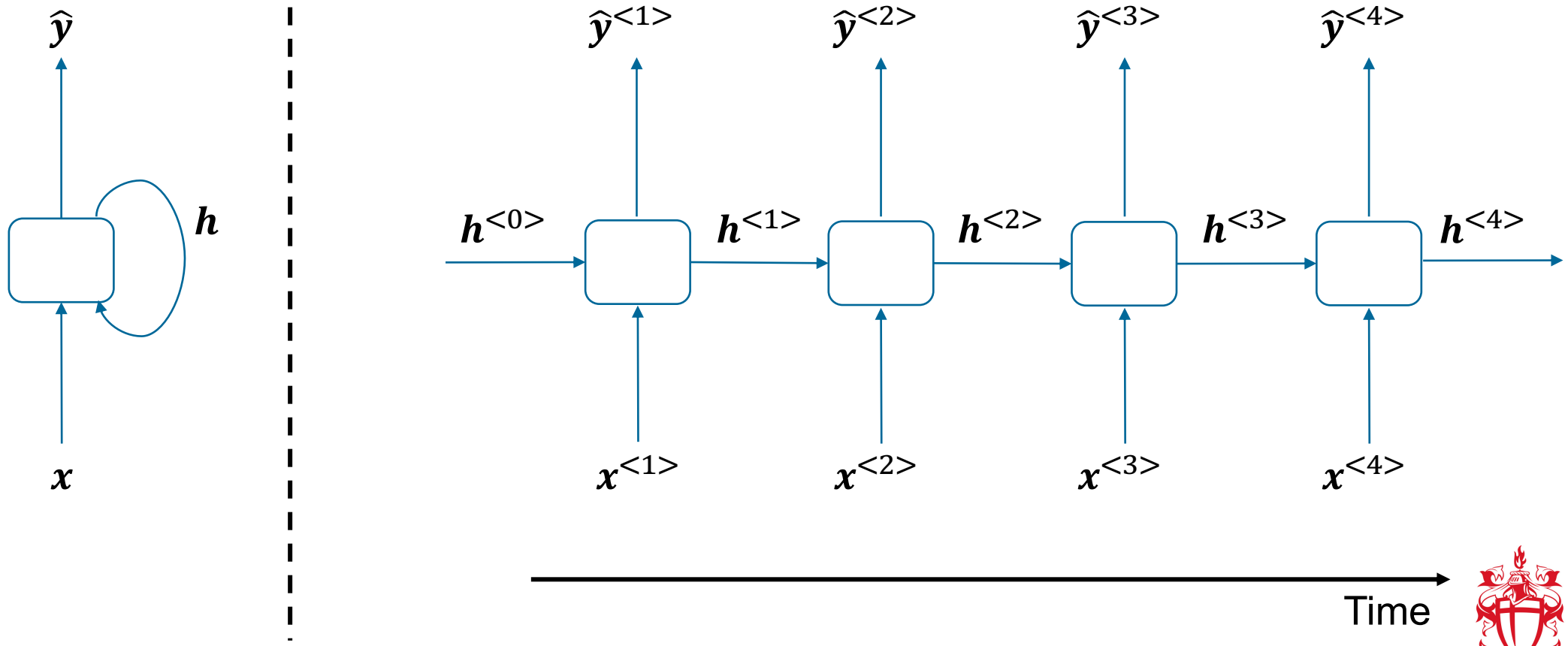
- We discuss attention mechanisms and how they can improve RNNs
- We then introduce transformers, which rely solely on attention and don’t use recurrence

Adding attention

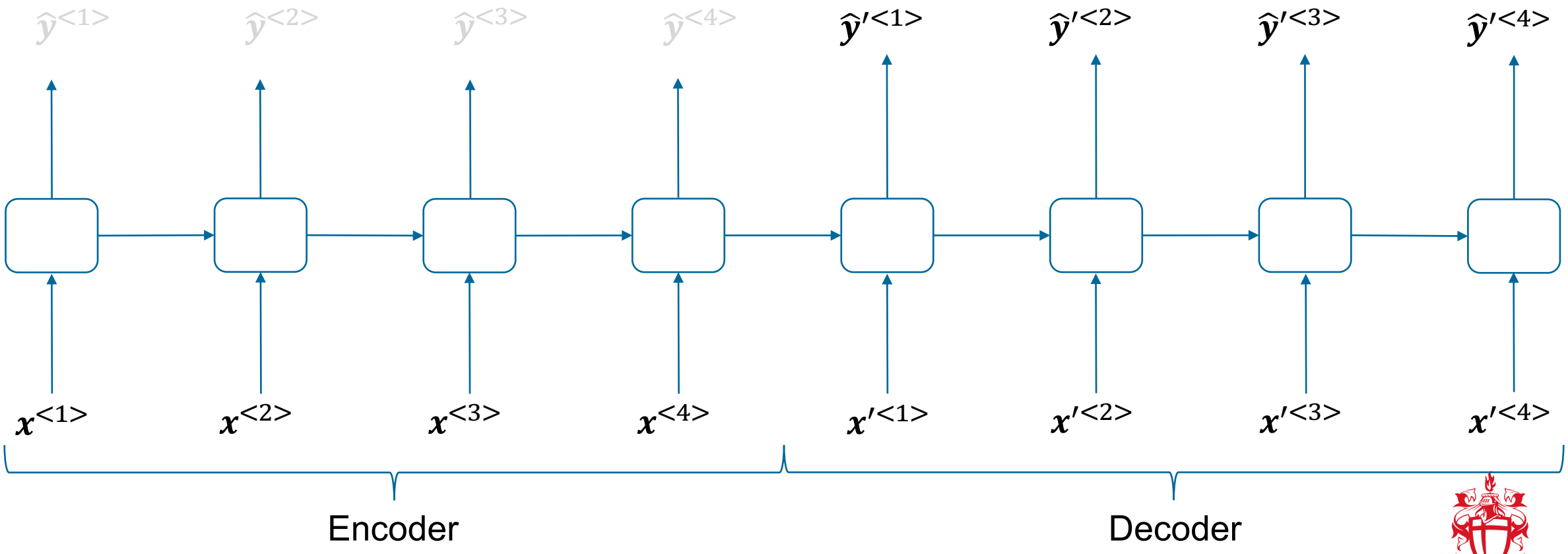
A brief summary of RNNs

- At each time step, take the input and the “memory” (or *state*) from the previous time step to compute the output
- Use the same parameters (and, also, activation functions) across different time steps

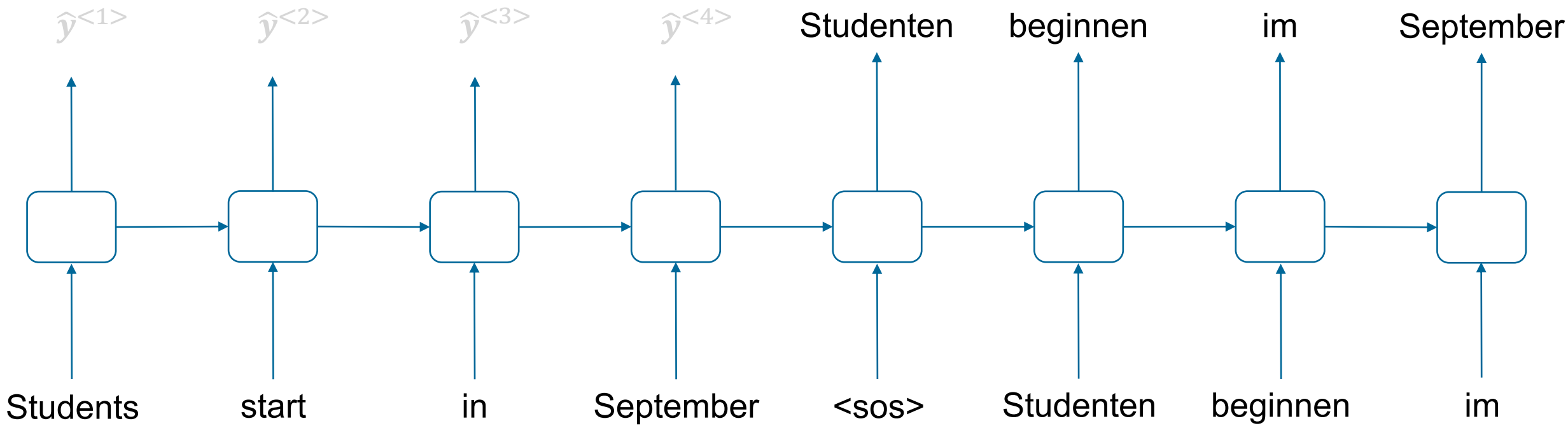
A typical RNN



Encoder-decoder networks

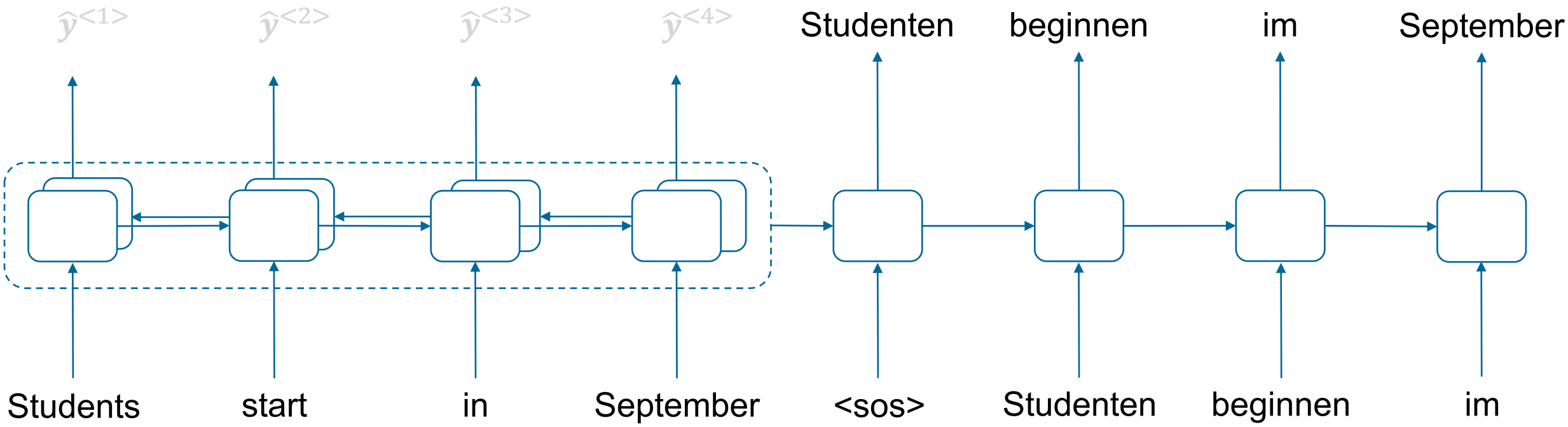


Encoder-decoder networks, for example in translation



BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Encoder-decoder networks, adding in a bidirectional RNN

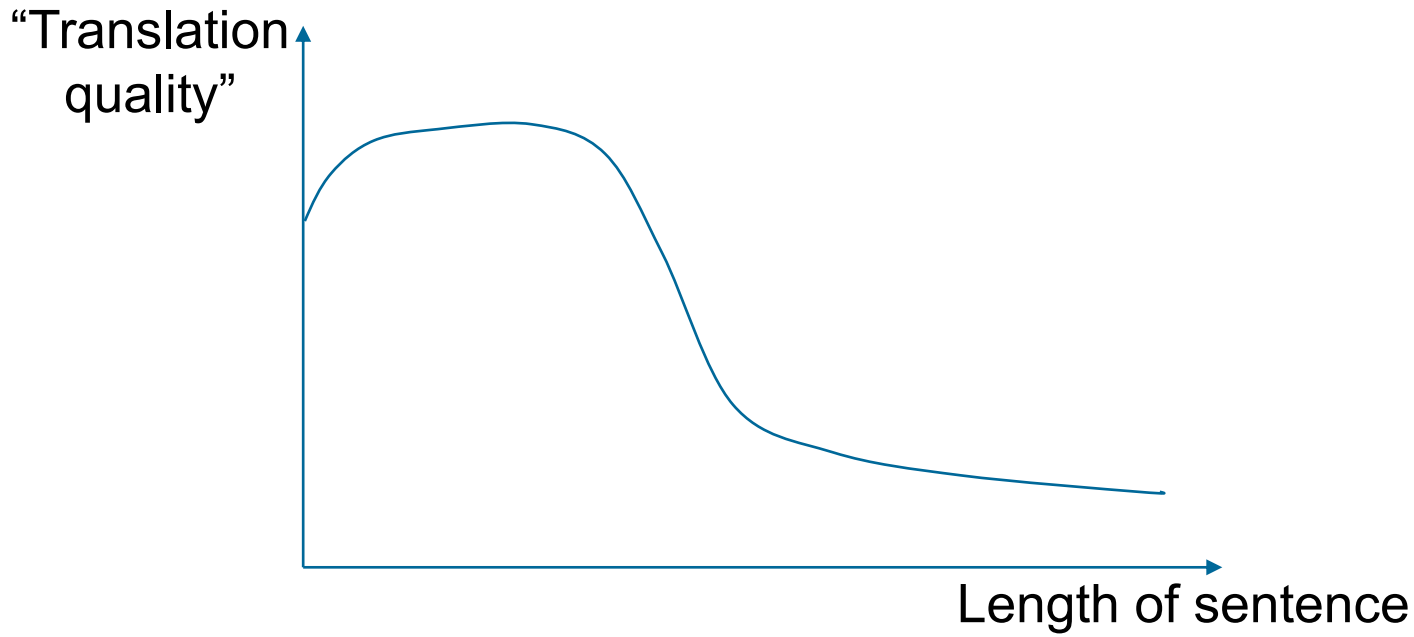


BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Long sentences still pose a problem

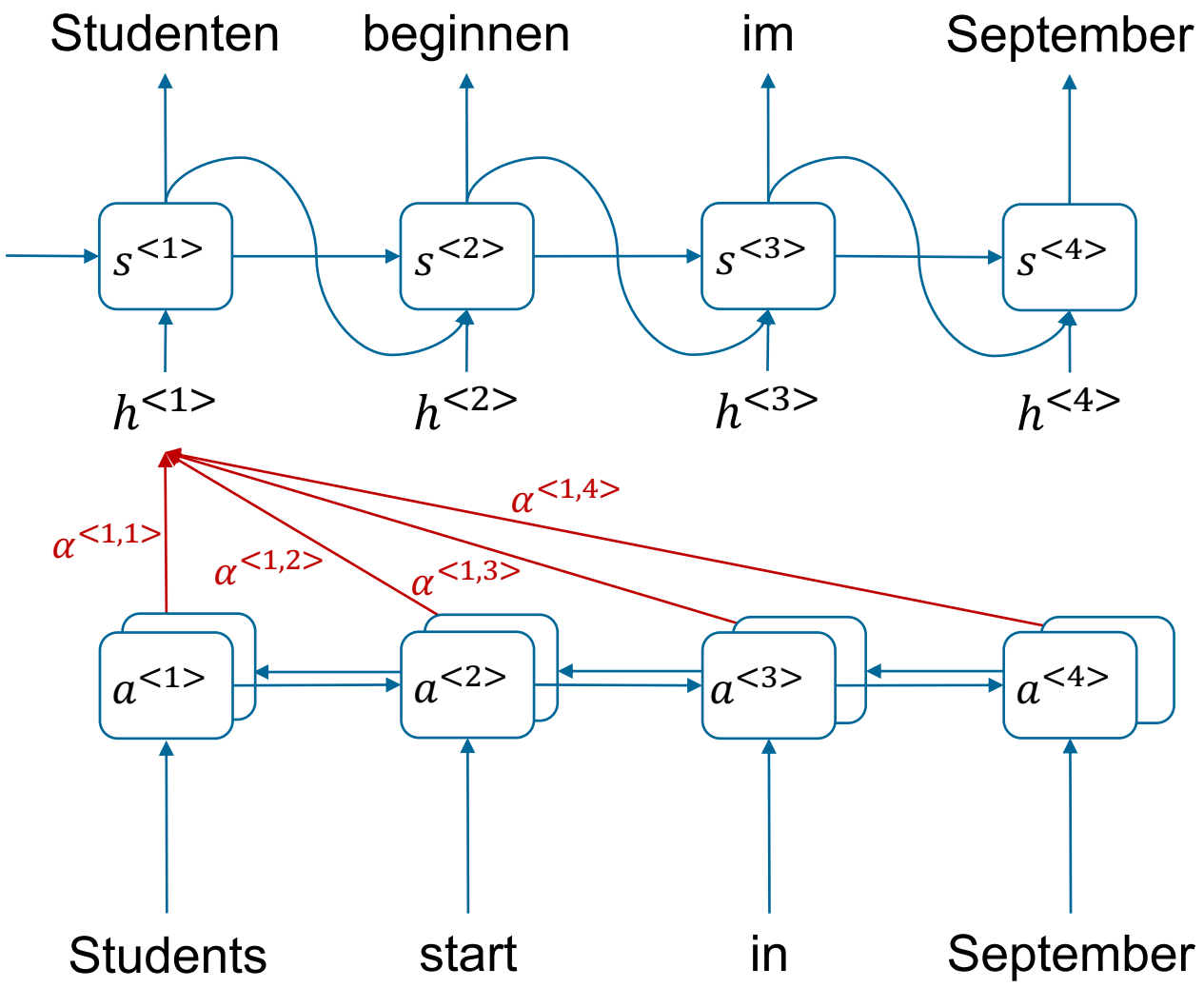
Students start in September, have three terms with classes, finish a project, and successfully complete their degree the following summer.

Die Studenten beginnen im September, haben drei Semester Unterricht, schließen ein Projekt ab und schließen ihr Studium im folgenden Sommer erfolgreich ab.



BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Putting attention on parts of the input

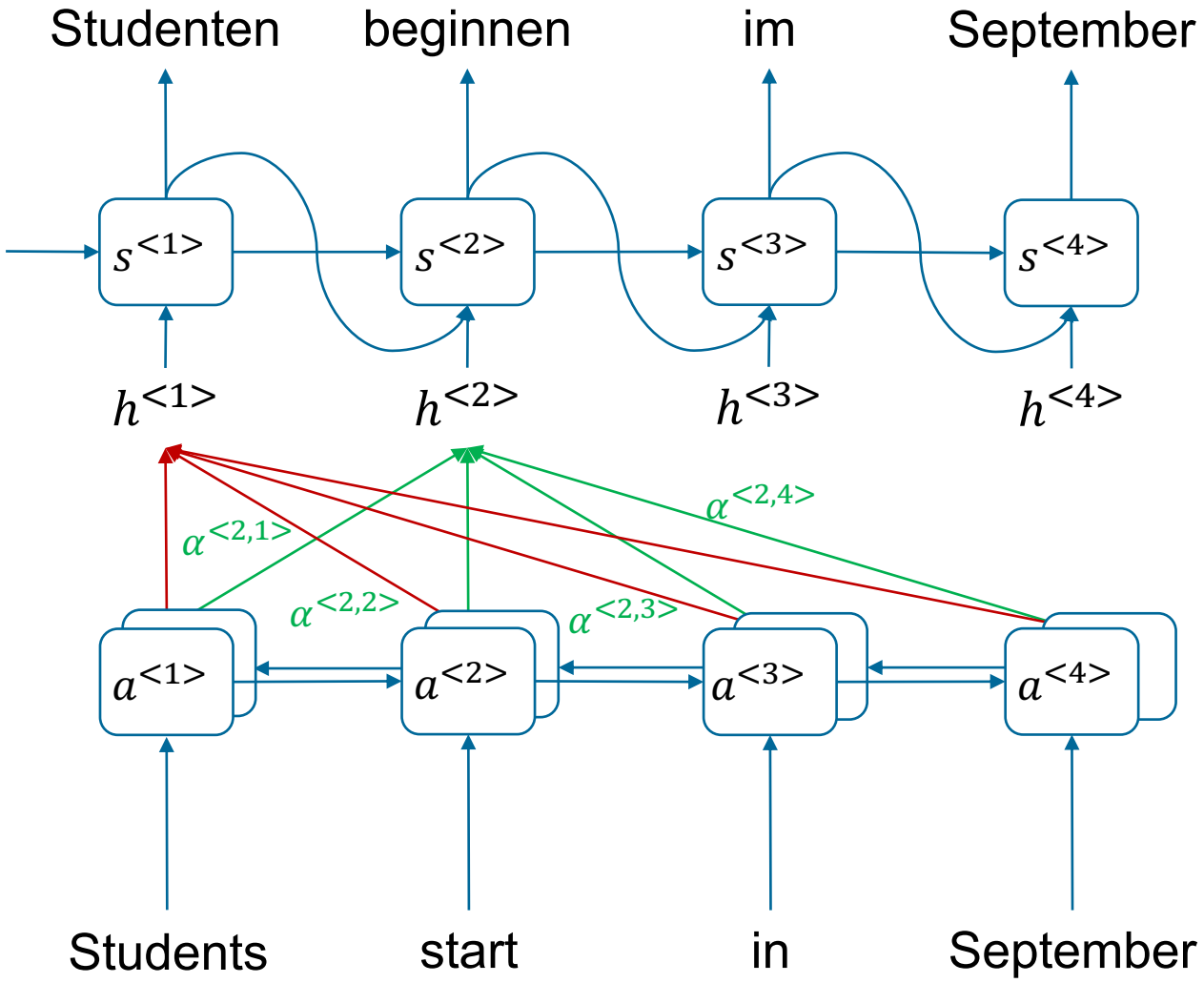


$h^{<1>} = \sum_t \alpha^{<1,t>} a^{<t>} = \alpha^{<1,1>} a^{<1>} + \alpha^{<1,2>} a^{<2>} + \alpha^{<1,3>} a^{<3>} + \alpha^{<1,4>} a^{<4>}$

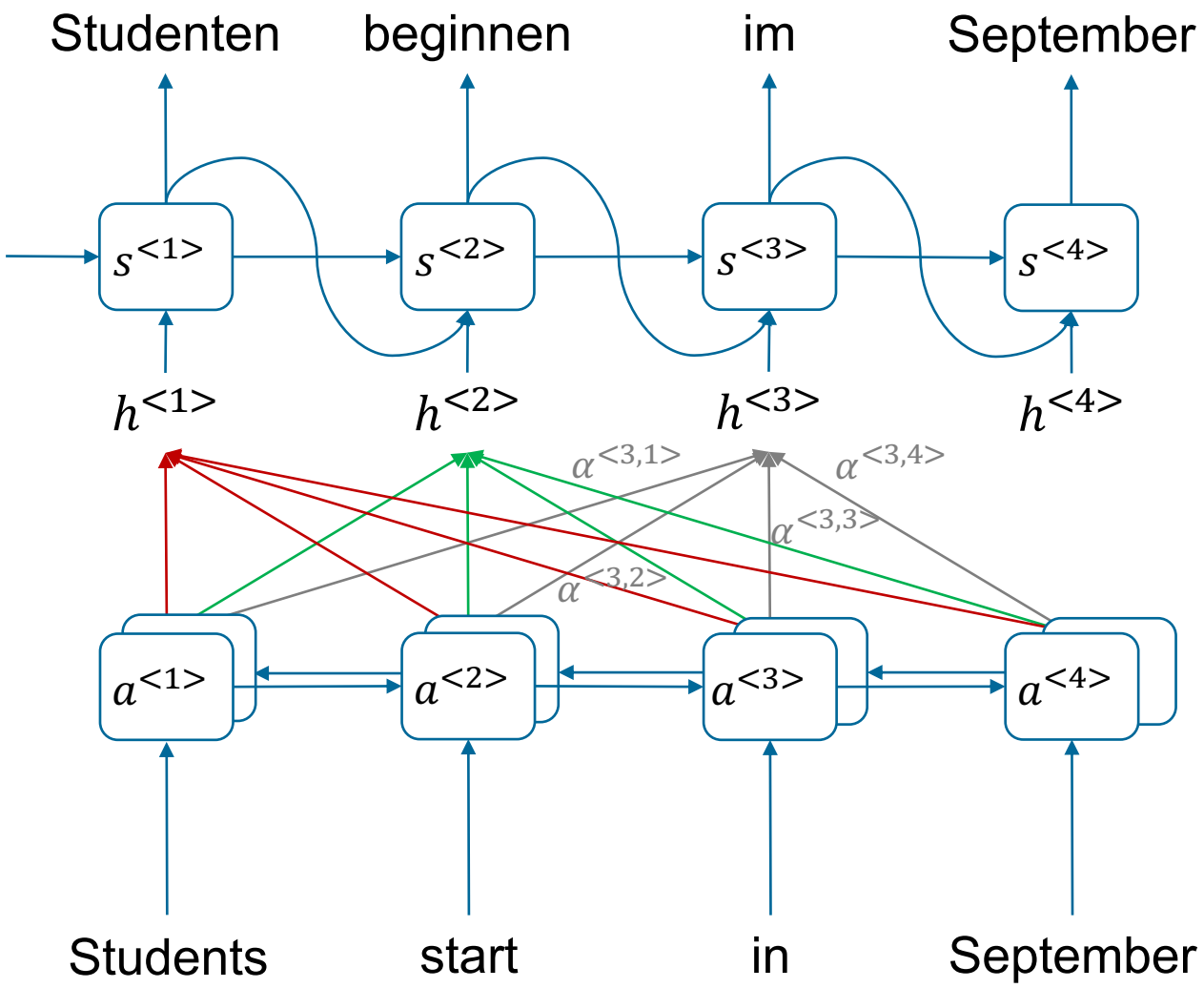
$\sum_t \alpha^{<1,t>} = 1$

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

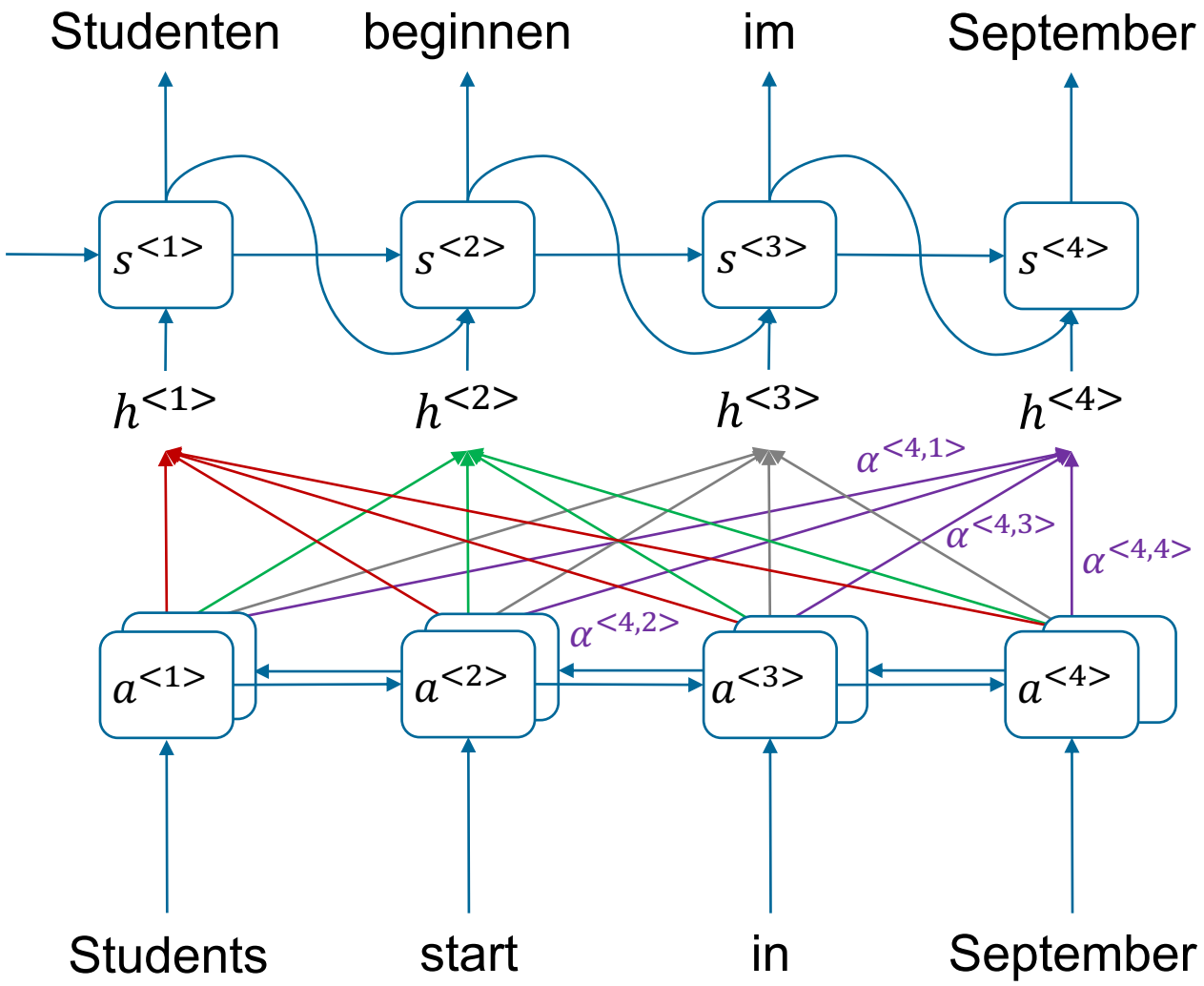
Putting attention on parts of the input



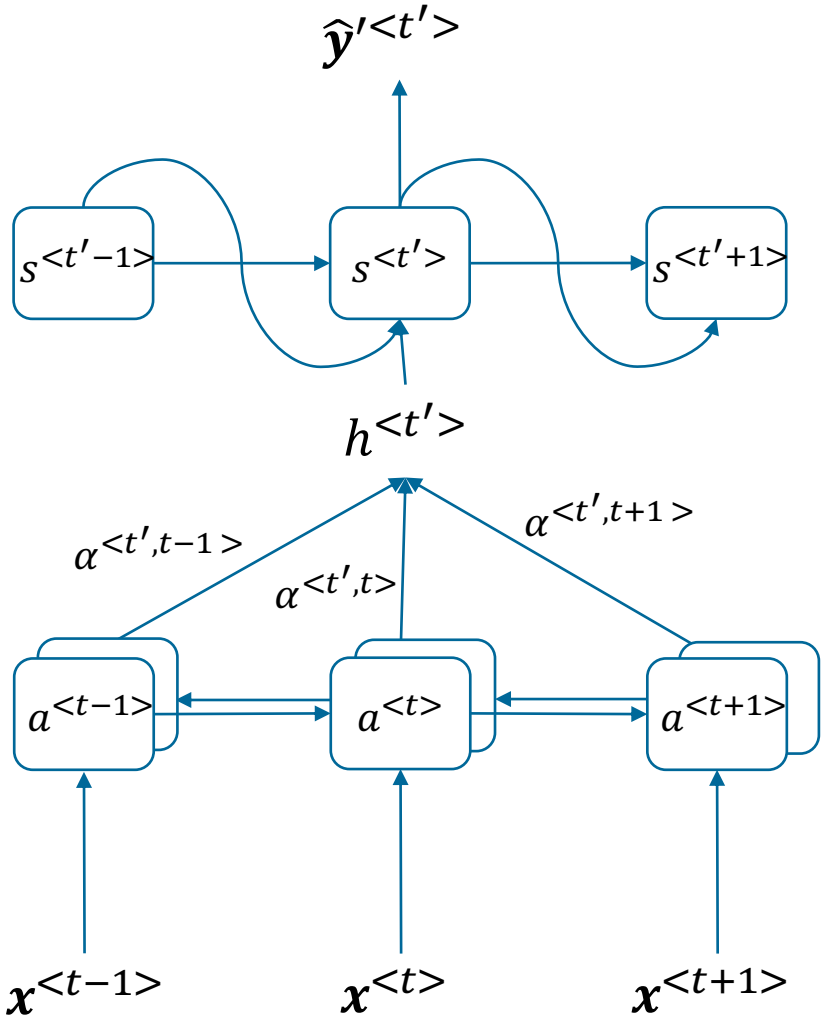
Putting attention on parts of the input

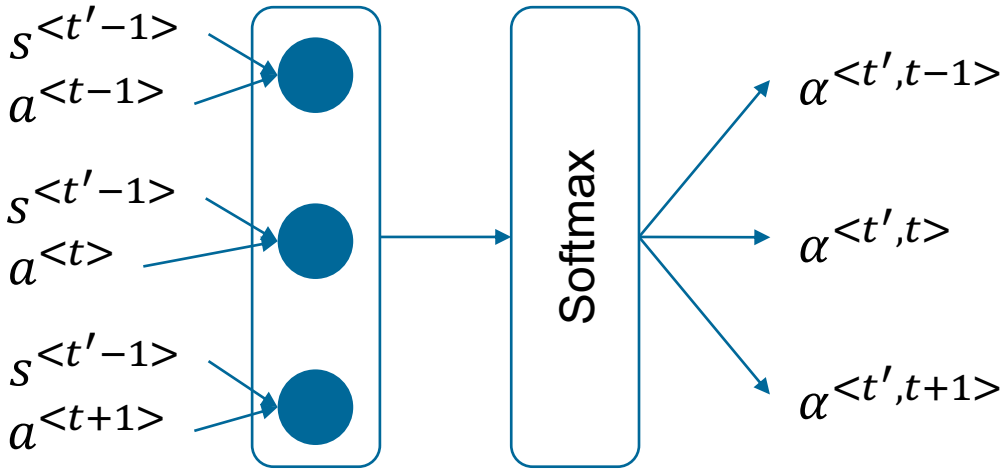


Putting attention on parts of the input



Computing attention





BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Attention is all you need

An overview

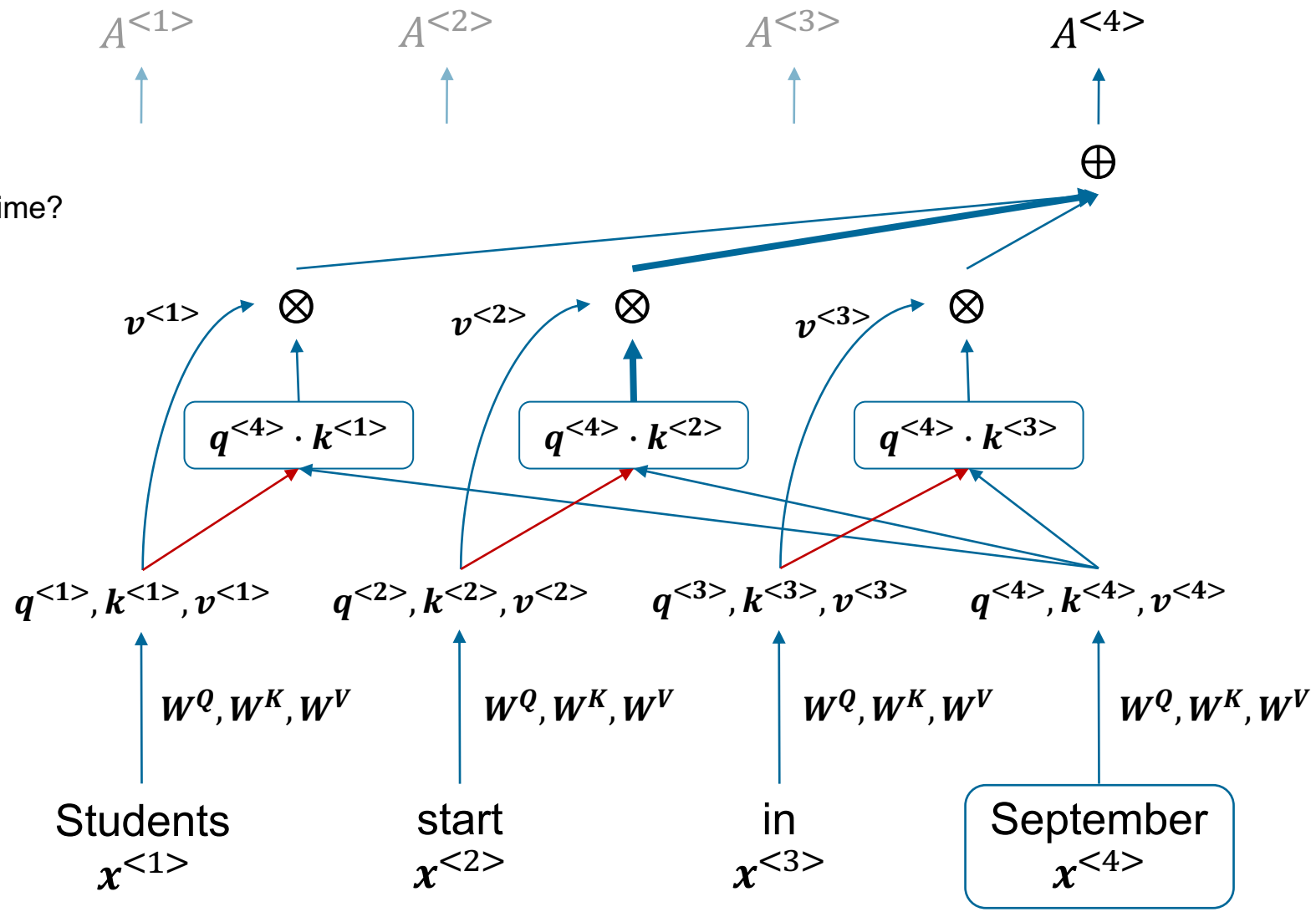
- 2017 paper by researchers at Google – probably most groundbreaking paper in last decade of machine learning
  - If you want to read one research paper only, it should be this
- Use attention exclusively (and add in some sequence information in another way)
- Motivation: recurrence cannot be parallelized by its very nature, drastically limiting NLP models
  - While motivation and main application in NLP, can analyze all sorts of data

Self-attention

We say $A(Q, K, V)$ is the attention-based representation of a word

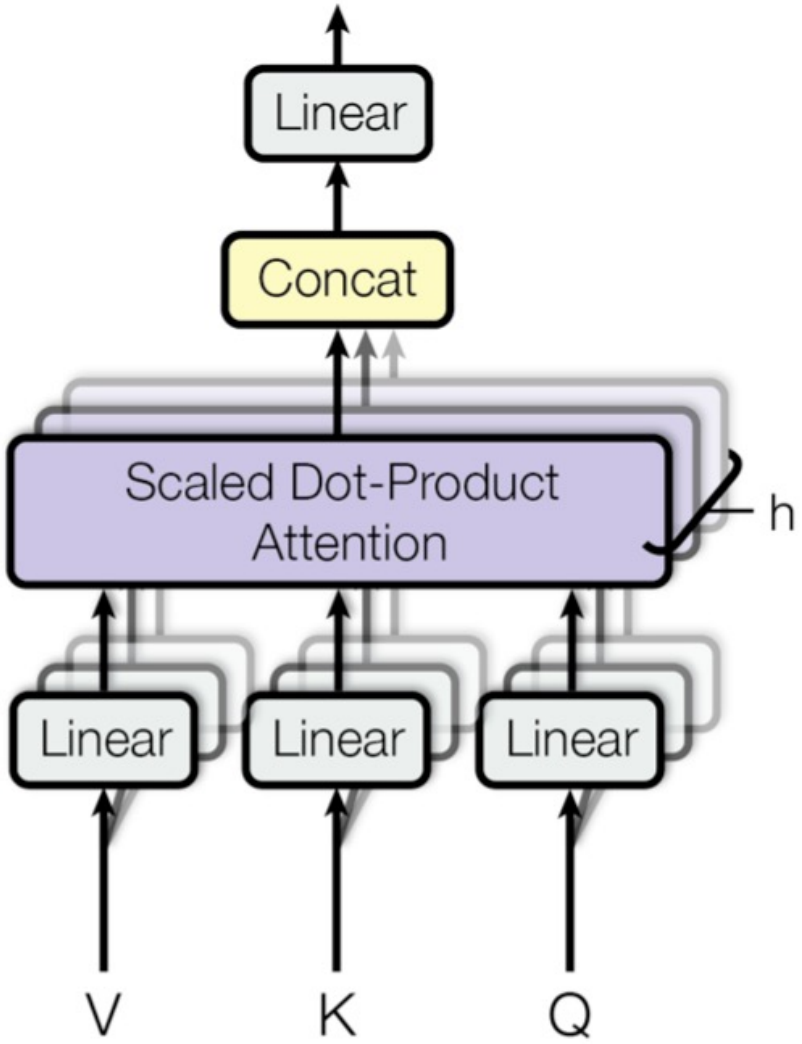
Query: What happens at this time?

$\boldsymbol{k}^{<1>}$: Persons
$\boldsymbol{k}^{<2>}$: Action
$\boldsymbol{k}^{<3>}$: Modifier



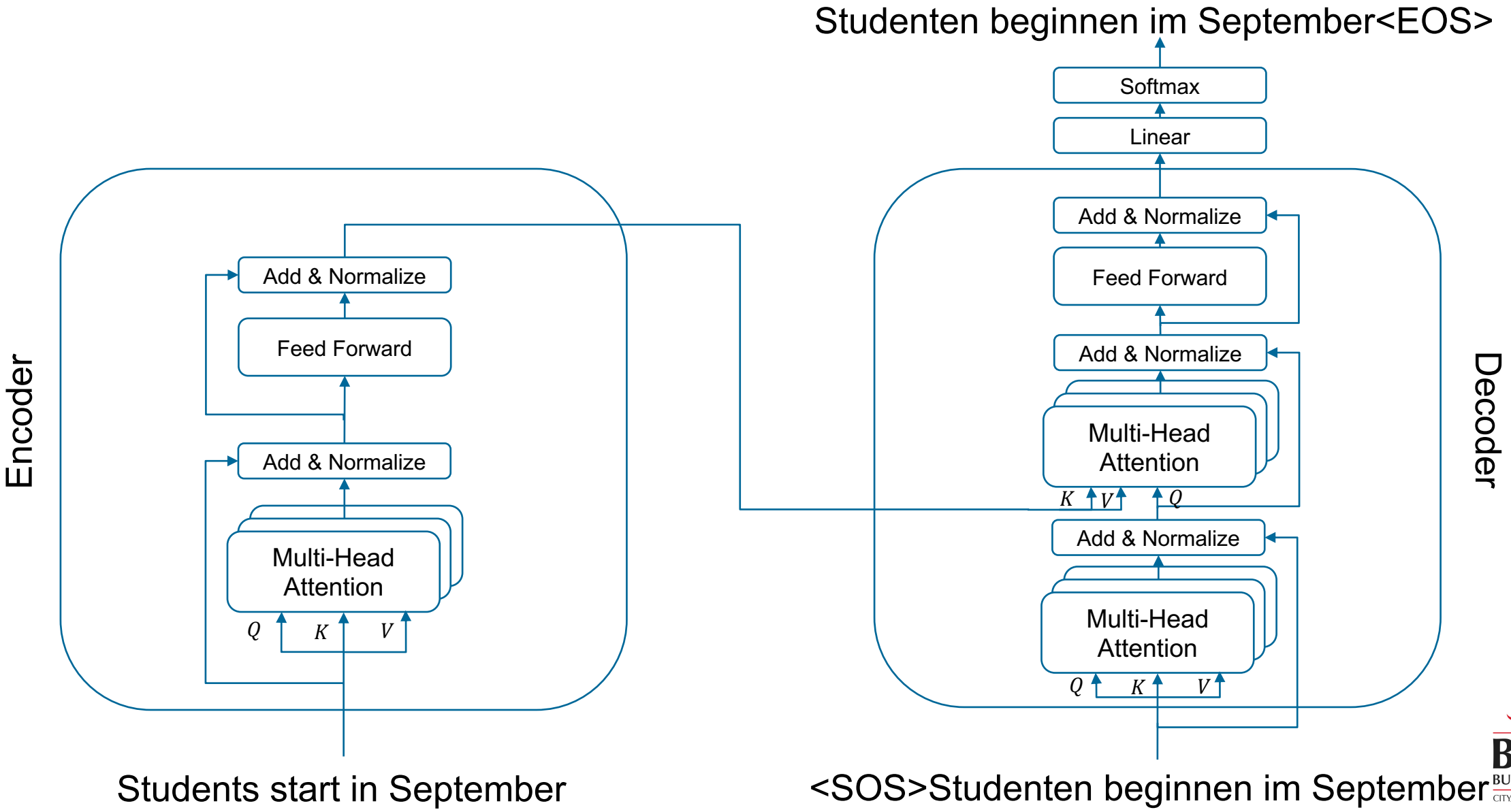
BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Multi-head attention layer



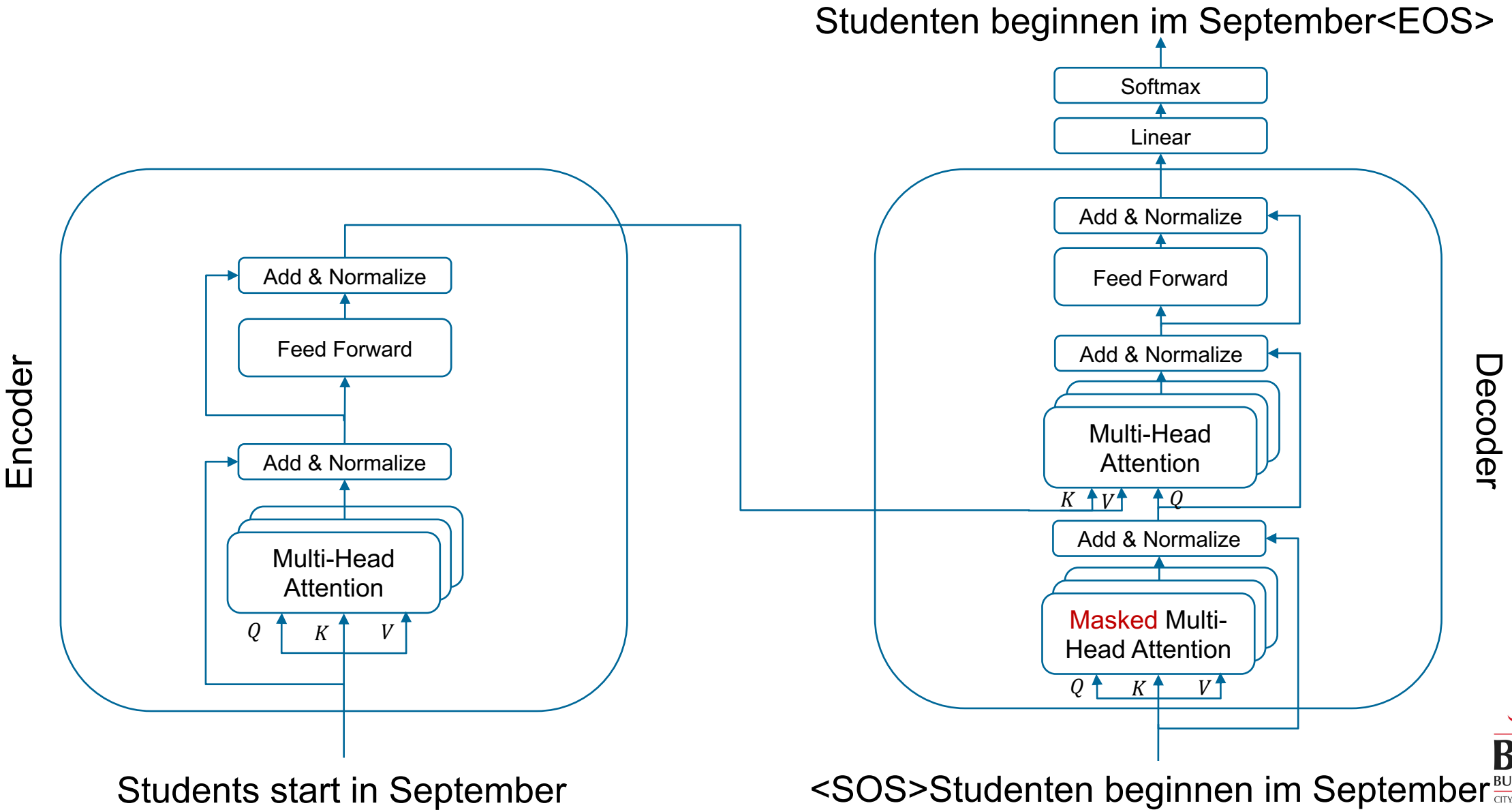
Source: Vaswani

Converting everything into the transformer model



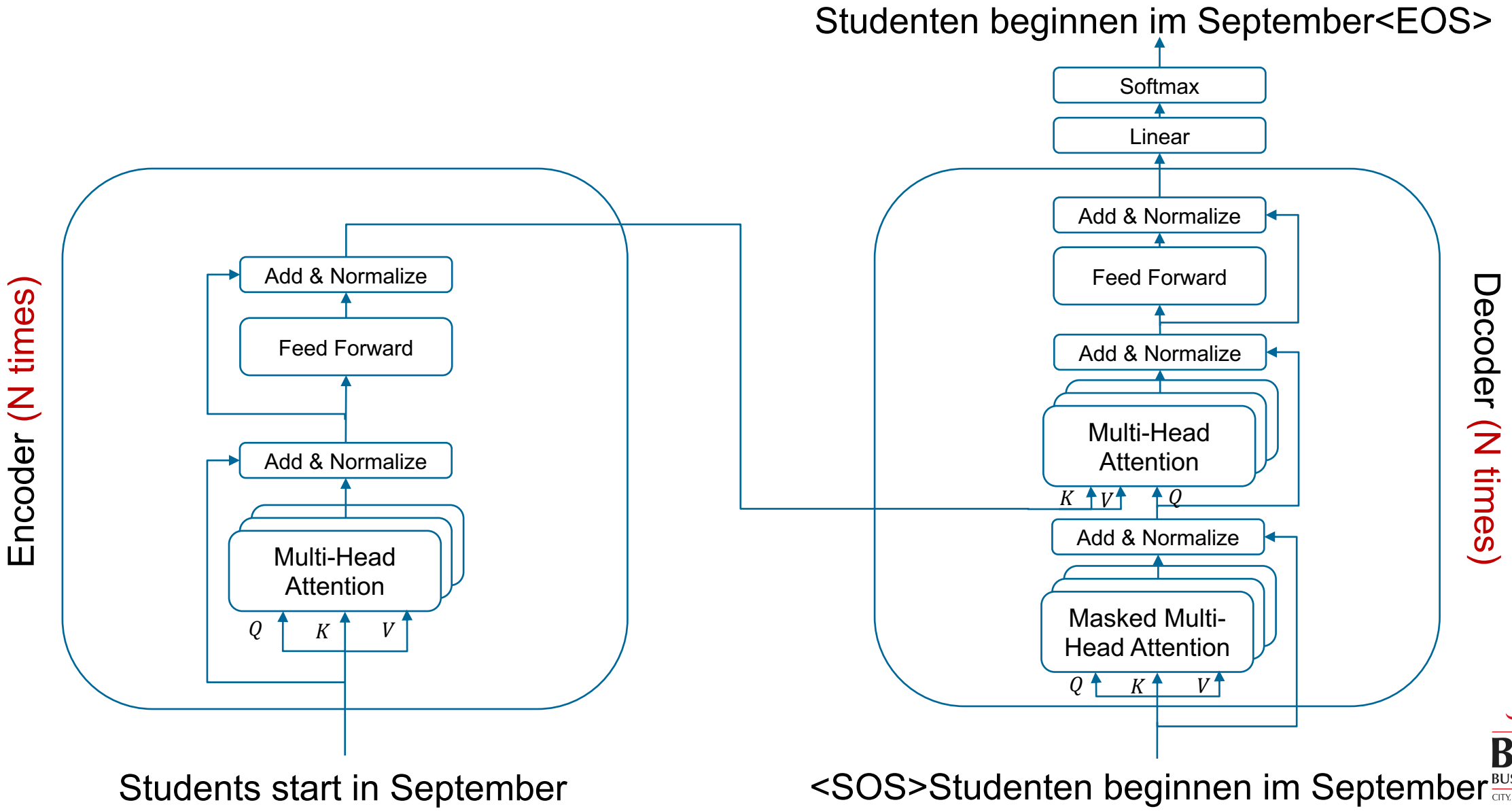
BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Converting everything into the transformer model



BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Converting everything into the transformer model



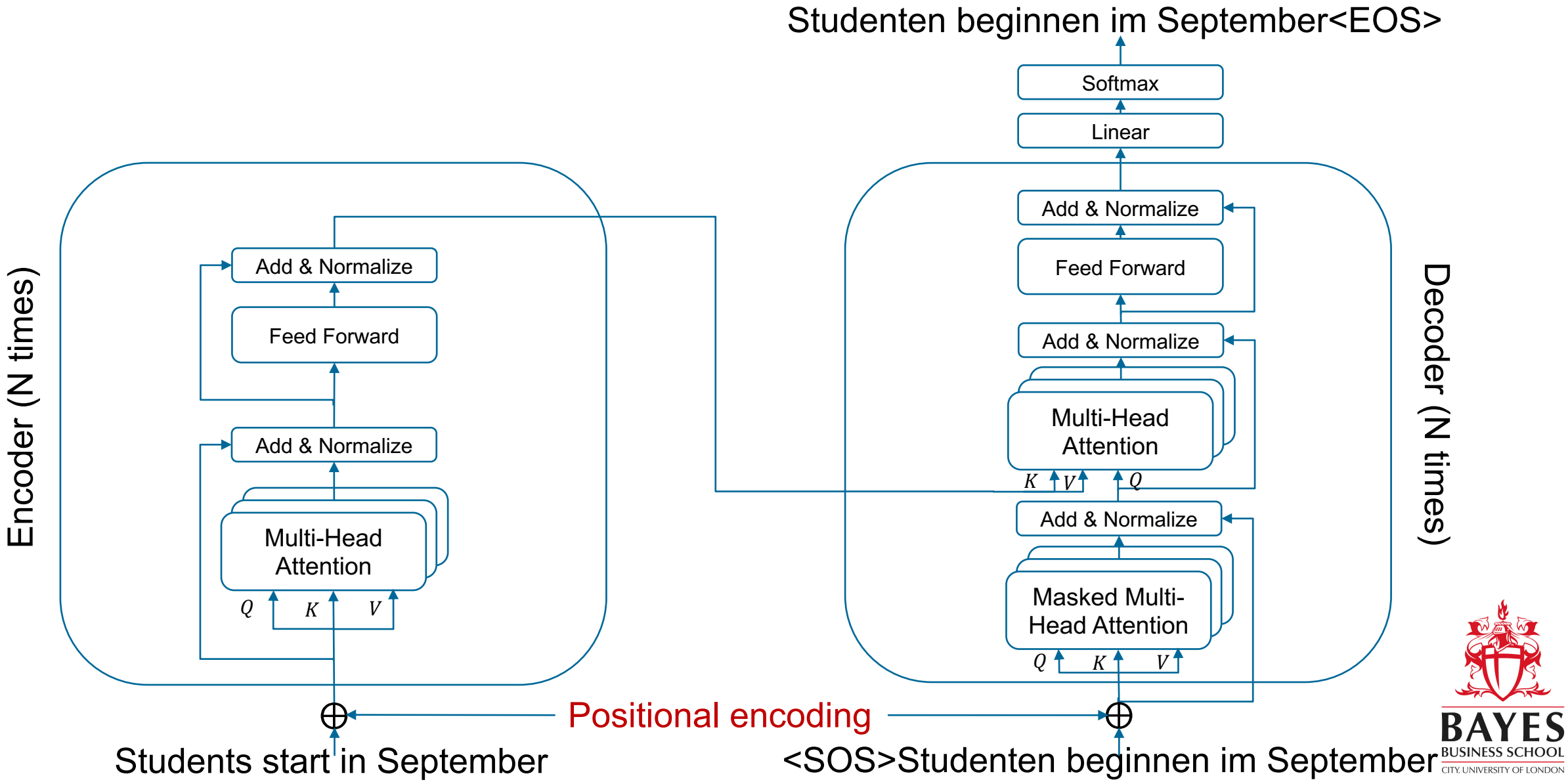
Why positions matter

Students start in September, have three terms with classes, finish a project, and successfully complete their degree the following summer.

Students start **the following summer**, have three terms with classes, finish a project, and successfully complete their degree **in September**.

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Converting everything into the transformer model



BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Positional encoding added to the word embedding – a first attempt

$$\text{September} \quad + \quad \text{Position encoding} \quad = \quad \text{New vector}$$

$$\begin{pmatrix} -0.21 \\ 0.45 \\ -0.84 \\ 0.13 \end{pmatrix} \qquad \begin{pmatrix} 4 \\ 4 \\ 4 \\ 4 \end{pmatrix} \qquad \begin{pmatrix} 3.79 \\ 4.45 \\ 3.16 \\ 4.13 \end{pmatrix}$$

Positional encoding added to the word embedding – a second attempt

September	+	Position encoding	=	New vector
$\begin{pmatrix} -0.21 \\ 0.45 \\ -0.84 \\ 0.13 \end{pmatrix}$		$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$		$\begin{pmatrix} 0.79 \\ 1.45 \\ 0.16 \\ 1.13 \end{pmatrix}$

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Positional encoding added to the word embedding – using sinusoids

$$PE_{pos,i} = \begin{cases} \sin\left(\frac{pos}{10000^{i/d}}\right), if\ i\ is\ even \\ \cos\left(\frac{pos}{10000^{(i-1)/d}}\right), if\ i\ is\ odd \end{cases}$$

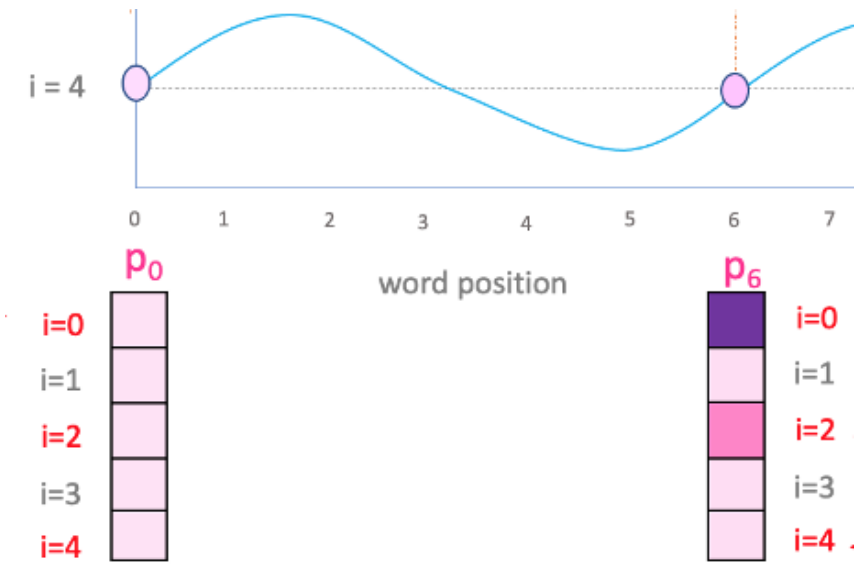
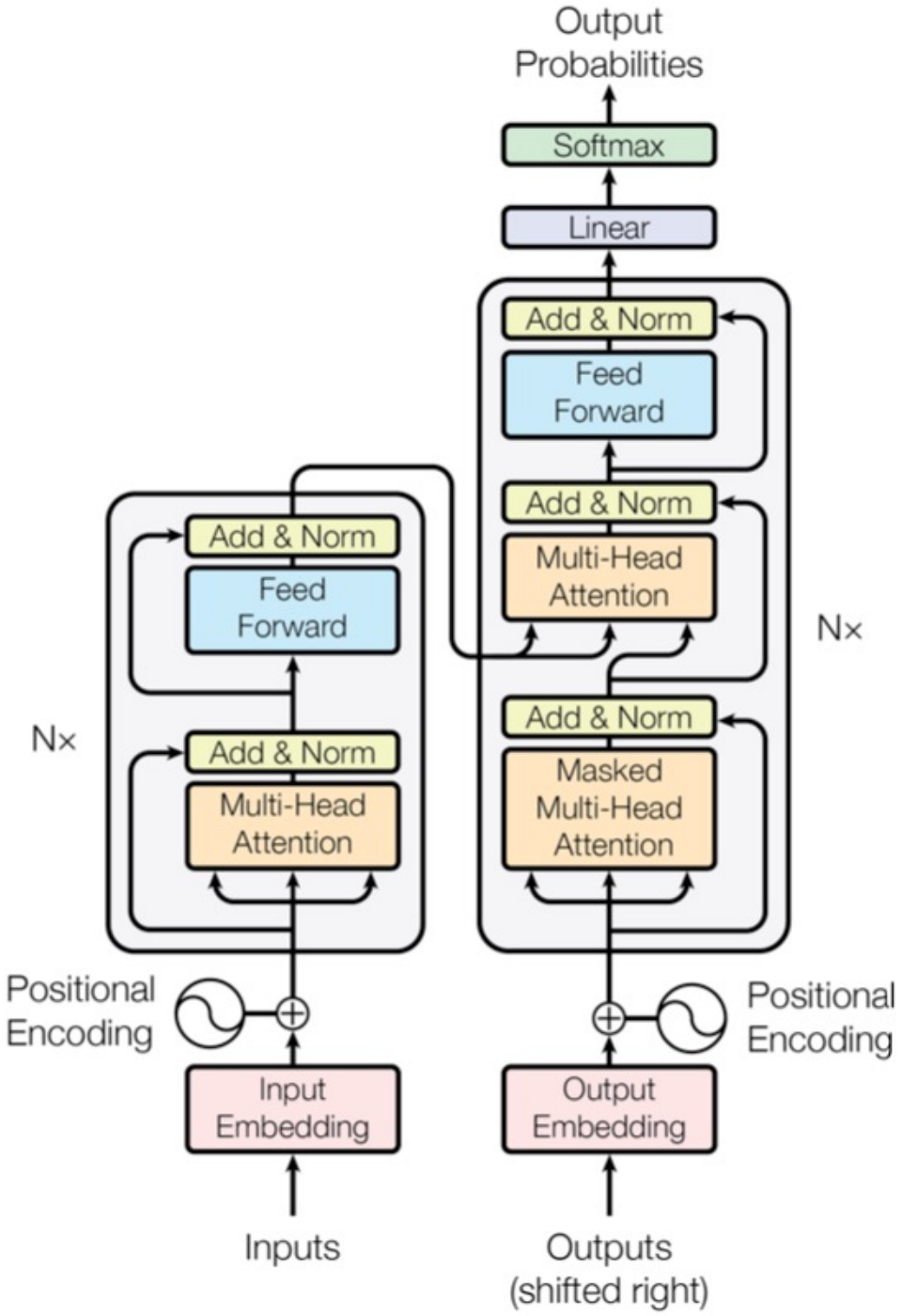


Image source: https://www.youtube.com/watch?v=dichIcUZfOw

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Putting it together



Source: Vaswani


See you in class!

Sources

- Bahdanau & Cho, 2015, Neural Machine Translation by Jointly Learning to Align and Translate: https://arxiv.org/pdf/1409.0473.pdf
- DeepLearning.AI, n.d.: deeplearning.ai
- Géron, 2019, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow
- Goodfellow, Bengio, Courville, 2016, The Deep Learning Book: http://www.deeplearningbook.org
- Vaswani et al., 2017, Attention is All You Need: https://arxiv.org/pdf/1706.03762.pdf