北京大学
PEKING UNIVERSITY

# Attention

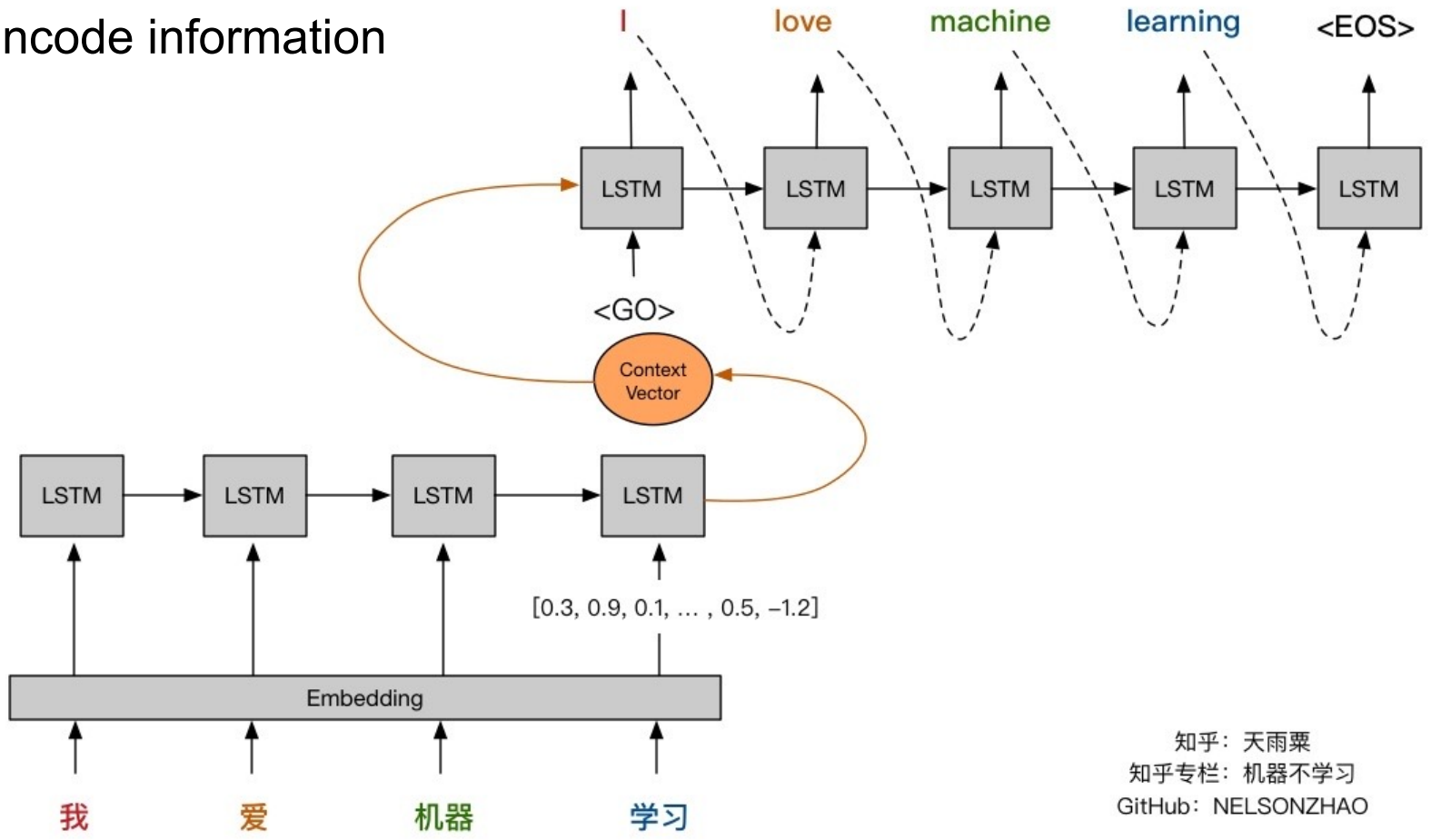
**Kun Yuan**

**Center for Machine Learning Research @ Peking University**

Oct. 24, 2023

# Traditional seq2seq model

RNN is hard to encode information that is far away




知乎：天雨粟
知乎专栏：机器不学习
GitHub：NELSONZHAO

# Attention

北京大学 PEKING UNIVERSITY

All you need is attention!

# Attention

How to capture the most valuable information from a pool of candidate?

Consider a pool of candidate information D = (k1,v1),(k2,v2),…,(km,vm)

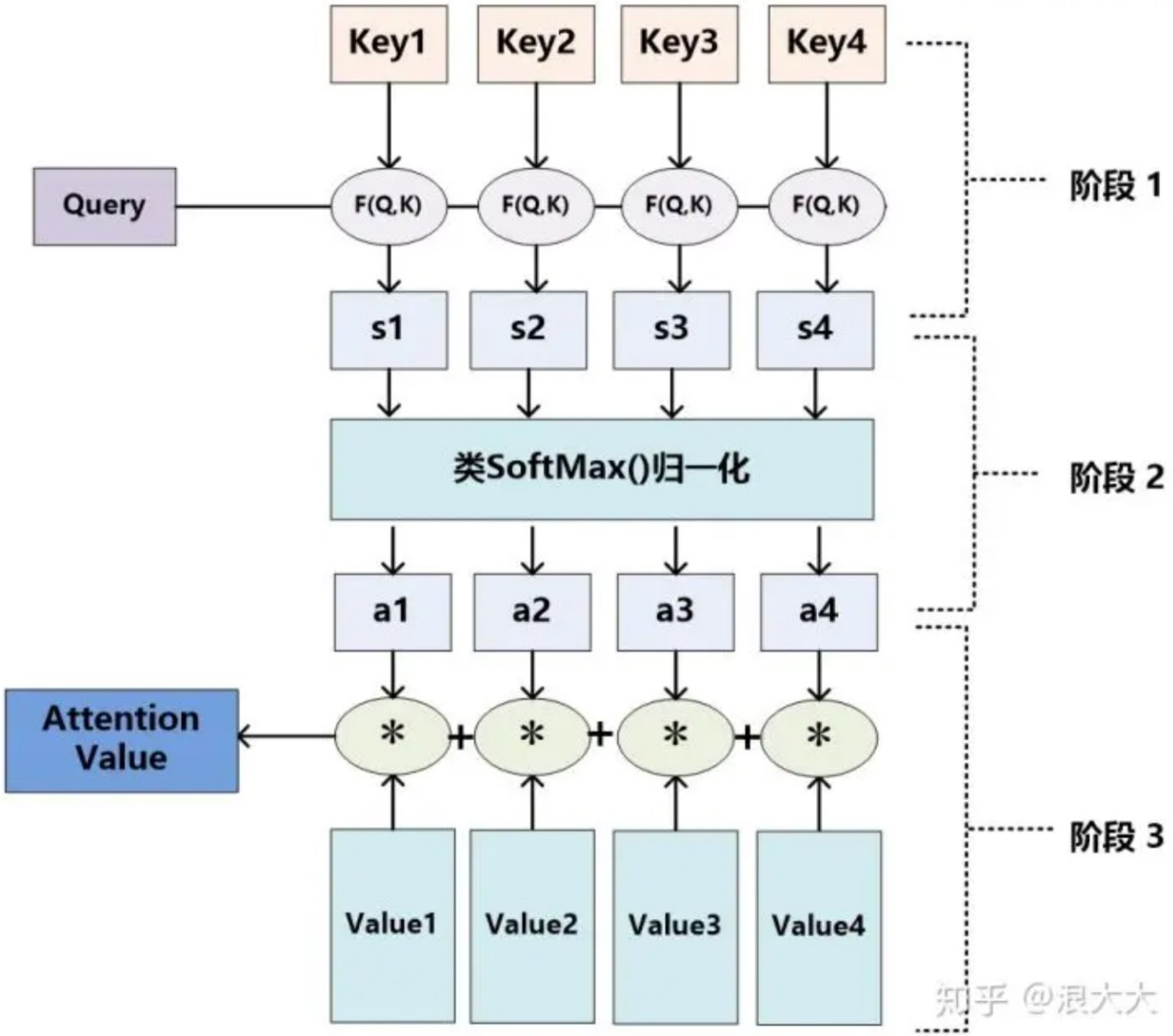
Given a query q, we can capture the most valuable information as follows

$$\text{Attention}(\mathbf{q}, \mathcal{D}) \stackrel{\text{def}}{=} \sum_{i=1}^{m} \alpha(\mathbf{q}, \mathbf{k}_i)\mathbf{v}_i,$$

where weight $\alpha$ is to evaluate how close the query q is to key ki

$$\alpha(\mathbf{q}, \mathbf{k}_i) = \text{softmax}(a(\mathbf{q}, \mathbf{k}_i)) = \frac{\exp(\mathbf{q}^\top \mathbf{k}_i/\sqrt{d})}{\sum_{j=1} \exp(\mathbf{q}^\top \mathbf{k}_j/\sqrt{d})}.$$

# Attention


Key1
Key2
Key3
Key4
Query
F(Q,K)
F(Q,K)
F(Q,K)
F(Q,K)
s1
s2
s3
s4
类SoftMax()归一化
a1
a2
a3
a4
Attention Value
*
*
*
*
Value1
Value2
Value3
Value4
阶段 1
阶段 2
阶段 3
知乎 @浪大大

# Seq2Seq with attention