Unveiling the Power of Attention: A Deep Dive into Attention Mechanisms

Bilal FAYE, Nicolas FLOQUET

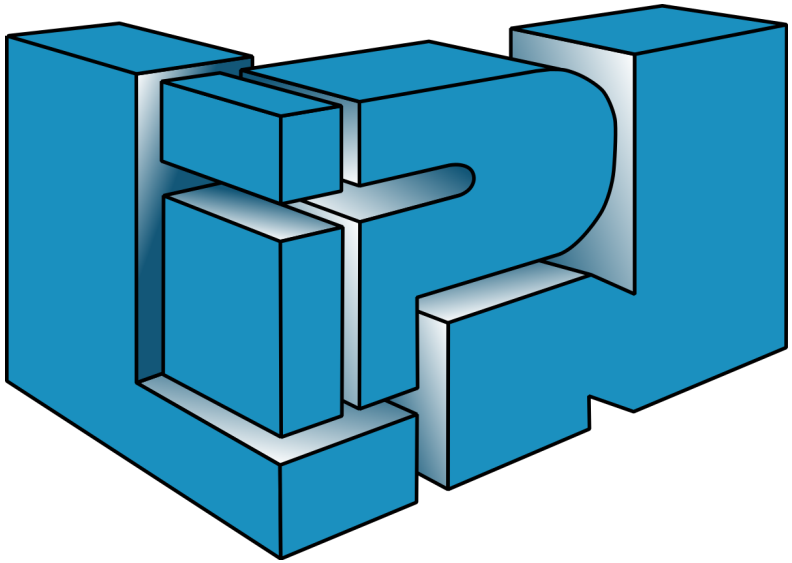


Table of contents

Introduction To Attention Mechanism

An attention mechanism in the context of machine learning refers to a computational model that allows a system to selectively focus on certain parts of input data while ignoring others.

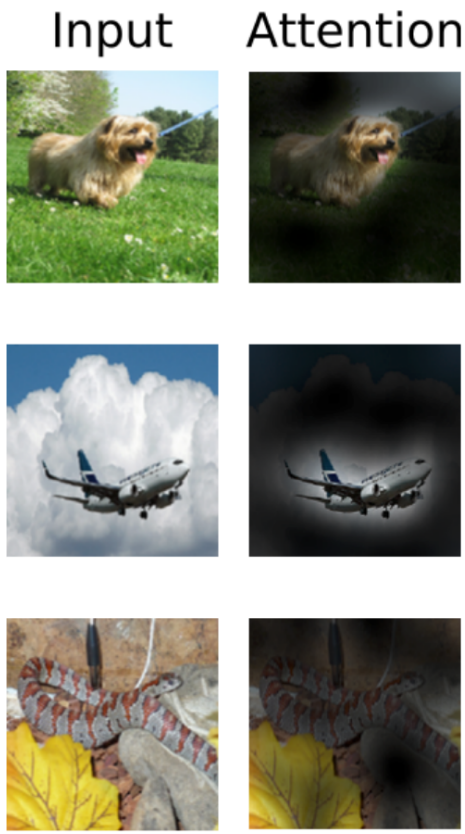


Figure – Attention Visualization on Image Classification [1]

1. Dosovitskiy et al., An Image is Worth 16x16 Words : Transformers for Image Recognition, 2020

Some key points highlighting the powerful aspects of attention mechanisms :

- Resource Allocation Efficiency
- Wide Range Applications
- Interpretable Neural Architectures
- Enhanced Performance in Specific Tasks
- Intuitive Explanations for Neural Behaviors

Broad Understanding of Attention Mechanisms

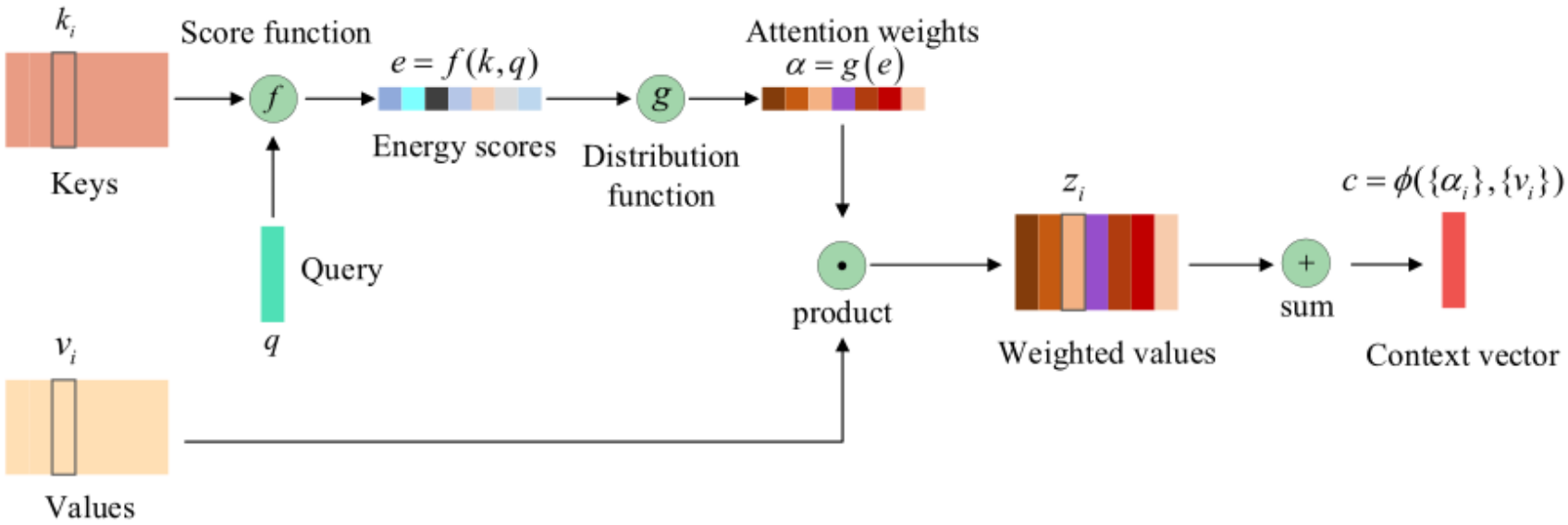


Figure – Unified Attention Model [2]

2. Zhaoyang et al., A review on the attention mechanism of deep learning, 2021

- **Keys** : This refers to the input data representations on which the model relies to identify relevant information or patterns in the sequence.
- **Query** : This represents what the model aims to search for or extract from the input data.
- **Values** : This corresponds to the actual information associated with each part of the input data (sequence).

The score function $\mathbf{e} = \mathbf{f}(\mathbf{k}, \mathbf{q})$ determines the matching or combination of keys and queries.

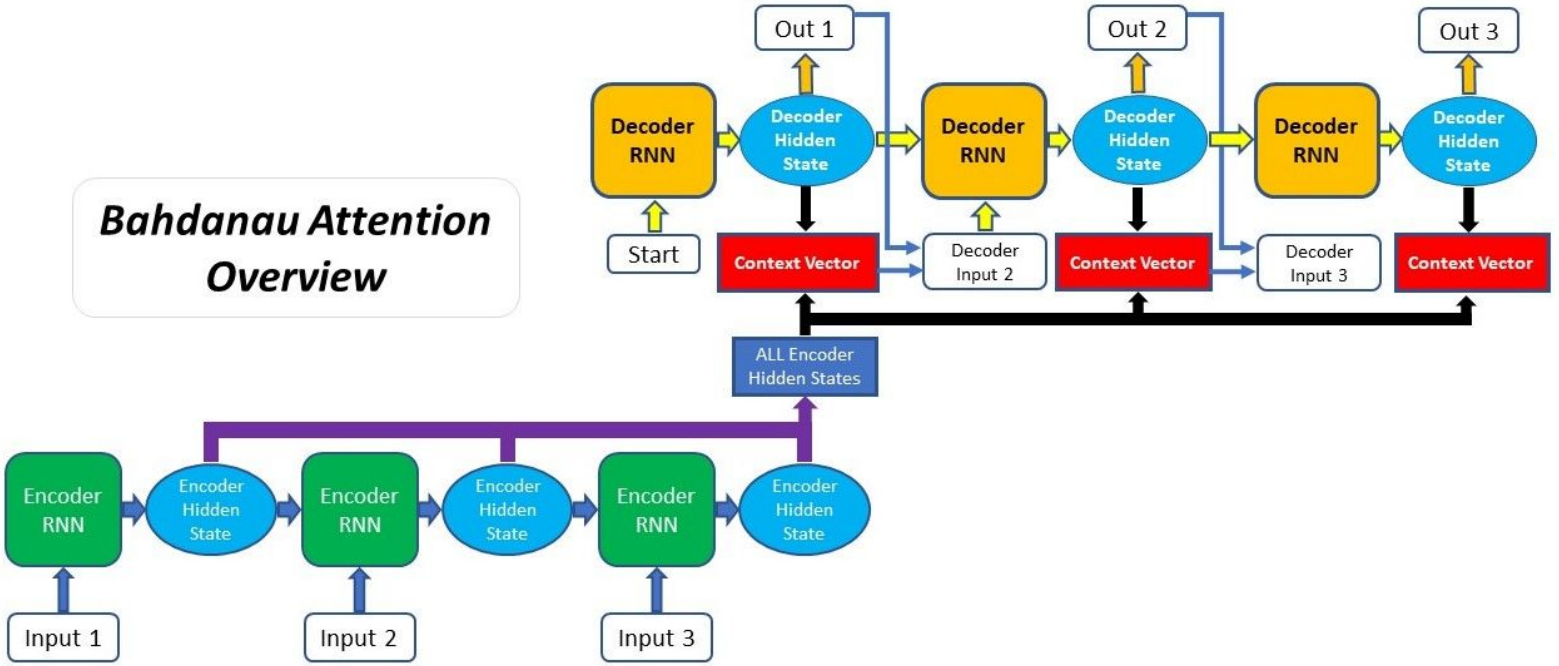
Additive attention [3] combines keys and queries through a summation operation :

Additive score function

$$\mathbf{f(q,k) = v^T activation(W_1 k + W_2 q)}$$

where $\mathbf{v}$, $\mathbf{W_1}$ and $\mathbf{W_2}$ are learnable parameters.

3. Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate, 2014


Bahdanau Attention Overview
Out 1
Out 2
Out 3
Decoder RNN
Decoder Hidden State
Decoder RNN
Decoder Hidden State
Decoder RNN
Decoder Hidden State
Start
Context Vector
Decoder Input 2
Context Vector
Decoder Input 3
Context Vector
ALL Encoder Hidden States
Encoder RNN
Encoder Hidden State
Encoder RNN
Encoder Hidden State
Encoder RNN
Encoder Hidden State
Input 1
Input 2
Input 3

Broad Understanding of Attention Mechanisms

Multiplicative (dot-product) attention [4] computes the relevance between keys and queries by taking their dot product :

Multiplicative score function

$$\mathbf{f(q, k) = q^T k}$$

In WMT'15 English → German task [5], authors found that parameterized additive attention slightly outperformed multiplicative attention.

4. Luong et al., Effective Approaches to Attention-based Neural Machine Translation, 2015
5. Britz et al., Massive exploration of neural machine translation architectures, 2017

Broad Understanding of Attention Mechanisms

Scaled multiplicative (dot-product) attention [6] computes the relevance between keys and queries by taking their dot product :

Scaled multiplicative score function

$$\mathbf{f(q, k)} = \frac{\mathbf{q^T k}}{\sqrt{\mathbf{d_k}}}$$

where $\mathbf{d_k}$ is the dimension of keys.

For small values of $\mathbf{d_k}$, both mechanisms perform similarly, but additive attention outperforms multiplicative attention without scaling for larger $\mathbf{d_k}$.

6. Vaswani et al., Attnetion is all you need, 2017

General attention [7] extends the concept of multiplicative attention by introducing a learnable matrix parameter $\mathbf{W}$:

General score function

$$\mathbf{f(q, k) = q^T W k}$$

where $\mathbf{W}$ is a learnable parameter.

This approach is applicable to keys and queries with distinct representations.

7. Luong et al., Effective Approaches to Attention-based Neural Machine Translation, 2015

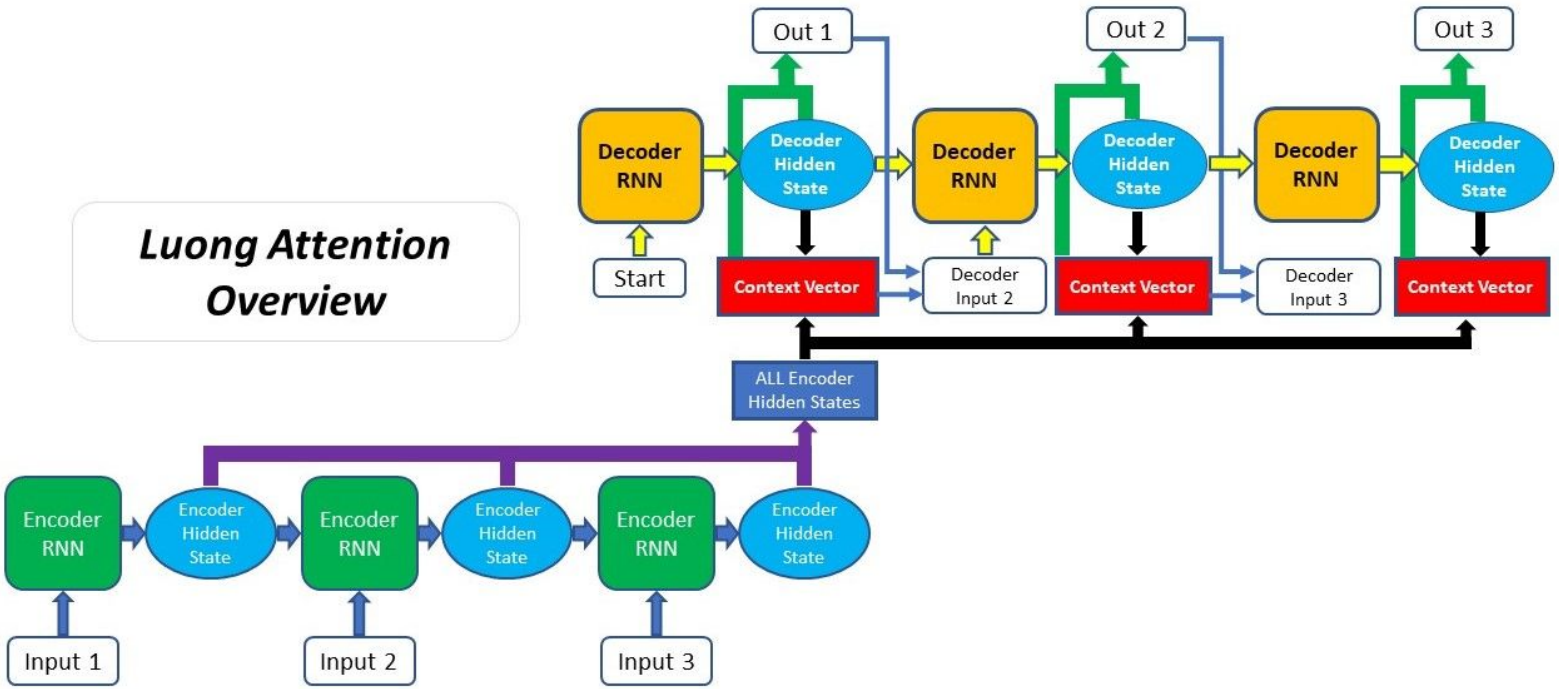
Concat attention [8] aims to derive the joint representation of the keys and queries instead of comparing them :

Concat score function

$$\mathbf{f(q, k) = v^T activation(W[k; q])}$$

where $\mathbf{W}$ is a learnable parameter.

8. Luong et al., Effective Approaches to Attention-based Neural Machine Translation, 2015


Luong Attention Overview
Out 1
Out 2
Out 3
Decoder RNN
Decoder Hidden State
Decoder RNN
Decoder Hidden State
Decoder RNN
Decoder Hidden State
Start
Context Vector
Decoder Input 2
Context Vector
Decoder Input 3
Context Vector
ALL Encoder Hidden States
Encoder RNN
Encoder Hidden State
Encoder RNN
Encoder Hidden State
Encoder RNN
Encoder Hidden State
Input 1
Input 2
Input 3

Location-based attention [9] are solely computed from the target hidden state :

Location-based score function

$$\mathbf{f}(\mathbf{q}, \mathbf{k}) = \mathbf{f}(\mathbf{q})$$

Energy scores ($\mathbf{f}$) depend solely on $\mathbf{q}$ rather than $\mathbf{K}$. Conversely, self-attention is calculated solely based on $\mathbf{K}$, without requiring $\mathbf{q}$.

9. Luong et al., Effective Approaches to Attention-based Neural Machine Translation, 2015

Similarity attention [10] compares the similarity between $\mathbf{K}$ and $\mathbf{q}$, which relied on cosine similarity. :

Similarity score function

$$\mathbf{f}(\mathbf{q}, \mathbf{k}) = \frac{\mathbf{q} \cdot \mathbf{k}}{\|\mathbf{q}\| \cdot \|\mathbf{k}\|}$$

Similarity attention is crucial in :

- semantic similarity assessments (in NLP)
- feature-based comparisons (in CV)

10. Graves et al., Neural turing machines, 2014

The distribution function **g** corresponds to the softmax, logistic or sigmoid, which normalize all the energy scores to a probability distribution.

After calculating attention weights and values, the context vector c is computed as follows :

Context vector

$$\mathbf{c} = \phi(\{\alpha_\mathbf{i}\}, \{\mathbf{v_i}\}),$$

where ϕ is a function that returns a single vector given the set of values and their corresponding weights.

$$\mathbf{z_i} = \alpha_\mathbf{i} \mathbf{v_i},$$

and

$$\mathbf{c} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{z_i},$$

where $\mathbf{z_i}$ is a weighted representation of an element in values and n is the dimension of $\mathbf{Z}$.

Categories of Attention Mechanisms

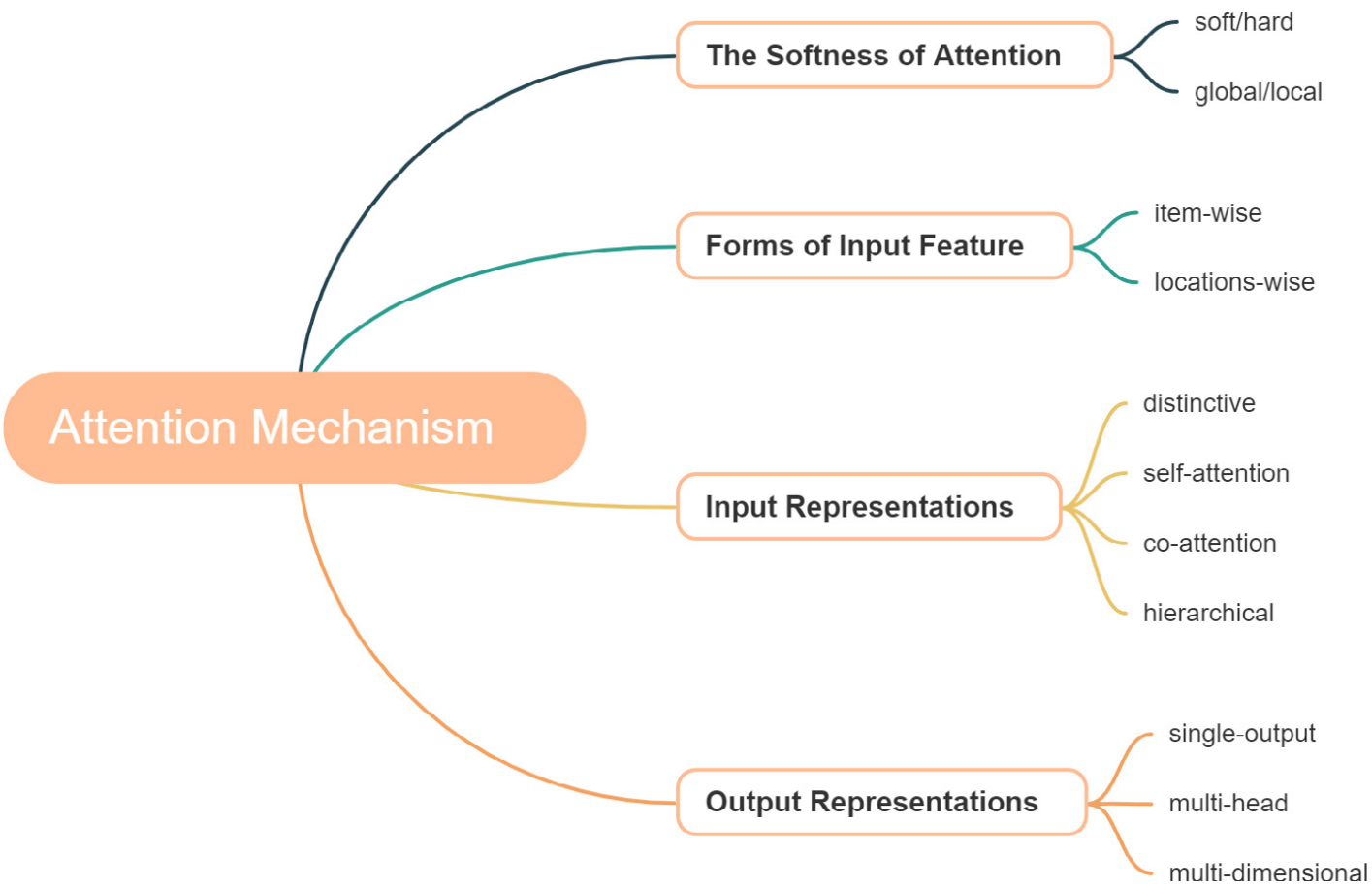


Figure – Unified Attention Model [11]

11. Zhaoyang et al., A review on the attention mechanism of deep learning, 2021

Categorie 1 : The Softness of Attention

- **Soft Attention :**
  - **Definition :** Soft (deterministic) attention calculates a context vector through a weighted average of all keys, facilitating differentiability with respect to inputs, thus enabling training via standard backpropagation methods.

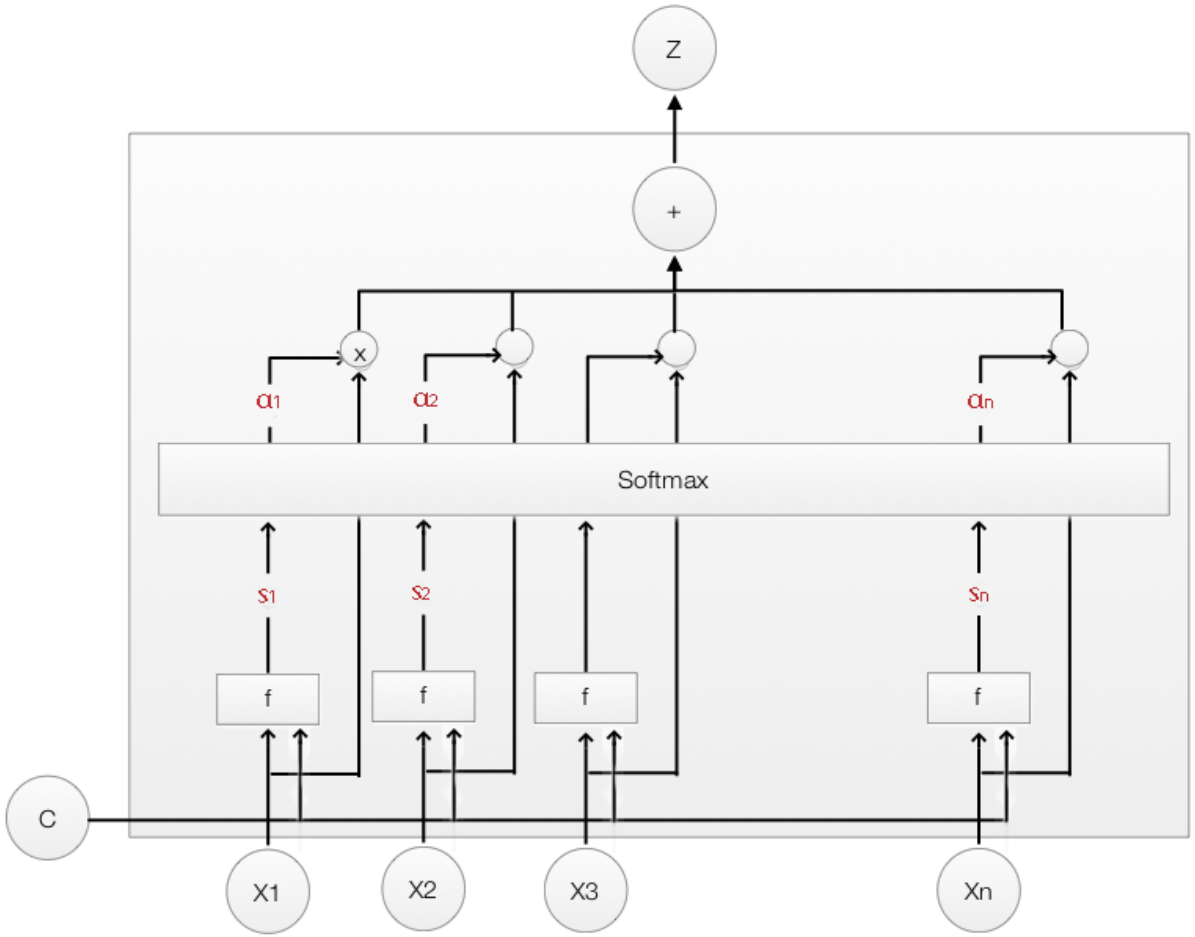


Figure – An instance demonstrating the application of soft attention.

- **Hard Attention :**
  - **Definition :** Hard (stochastic) attention is an attention mechanism that makes discrete decisions regarding which parts of the input sequence to focus on, resulting in non-differentiability with respect to inputs, thereby complicating training using standard backpropagation methods.

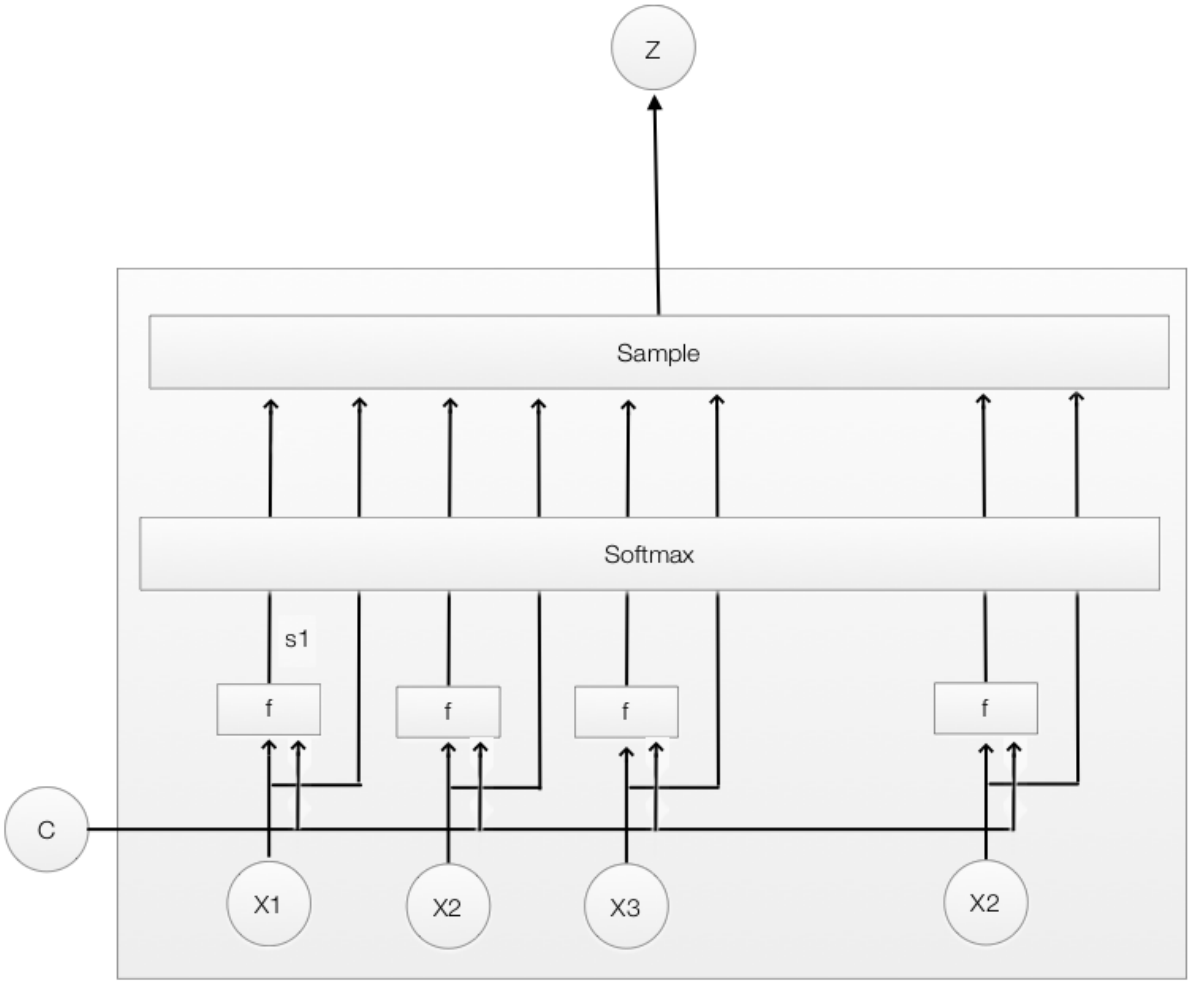
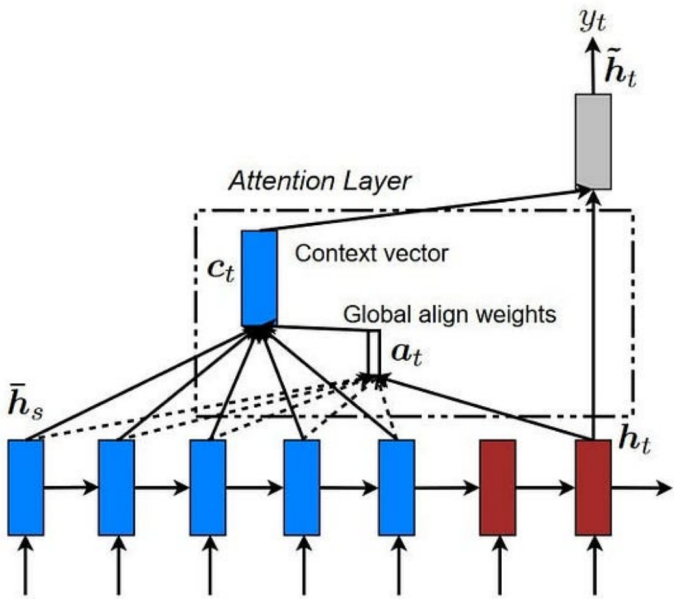
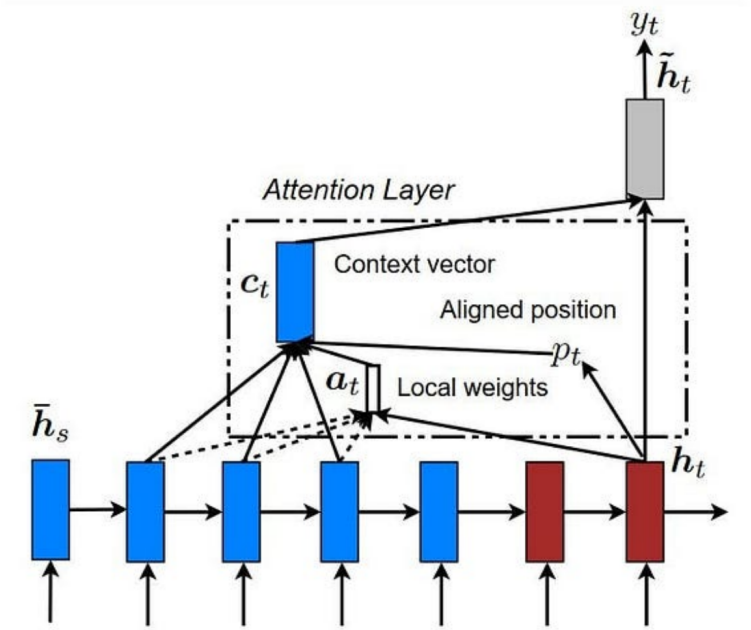


Figure – An instance demonstrating the application of hard attention.

- **Global Attention :**
  - **Definition :** Global attention is similar to soft attention, all source words are considered at a time.



- **Local Attention :**
  - **Definition :** In Local attention, only a subset of source words are considered at a time.



Categorie 2 : Form of Input Feature

- **Item-wise attention**
  - **Definition :** The item-wise attention requires that the input is either explicit items or an additional preprocessing (ex. word embeddings) step is added to generate a sequence of items (vectors) from the source data.
  - item-wise soft attention calculates a weight for each item, and then makes a linear combination of them.
  - Instead of a linear combination of all items, the item-wise hard attention stochastically picks one or some items based on their probabilities.

- **Location-wise attention**
  - **Definition :** location-wise attention is aimed at tasks that are difficult to obtain distinct input items (visual tasks).
  - The location-wise soft attention accepts an entire feature map as input and generates a transformed version through the attention module.
  - The location-wise hard attention stochastically picks a sub-region as input and the location of the sub-region to be picked is calculated by the attention module.

Categorie 3 : Input Representations

- **Distinctive Attention :**
  - **Definition :** Distinctive attention, as defined in the mentioned context, involves attention models with a single input and corresponding output sequence, where keys and queries are derived from two independent sequences.

- **Self-Attention :**
  - **Definition :** Self-attention is an attention mechanism in which each element in a sequence attends to all other elements in the same sequence.

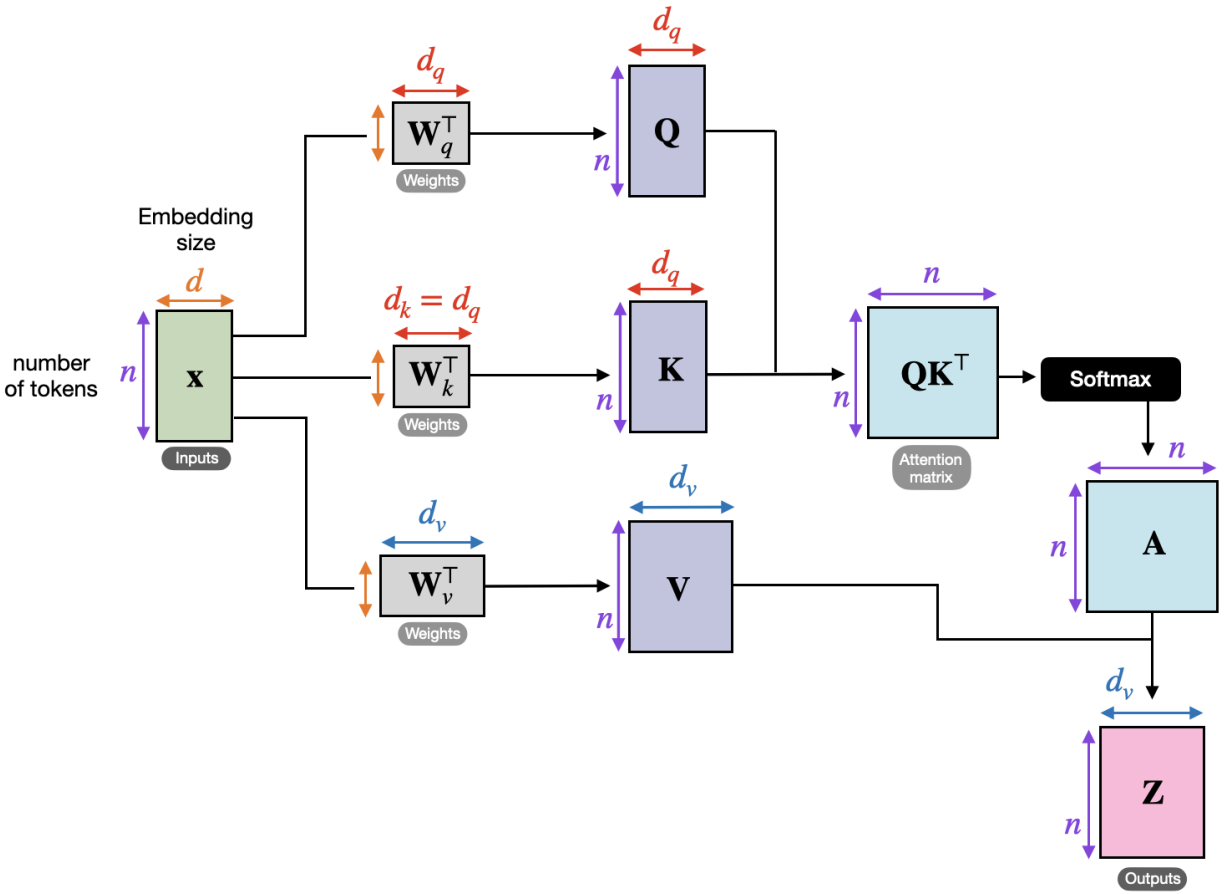


Figure – Application of self-attention

- **Cross-Attention :**
  - **Definition :** Cross-attention refers to scenarios where attention is applied between different parts of the input and output sequences.

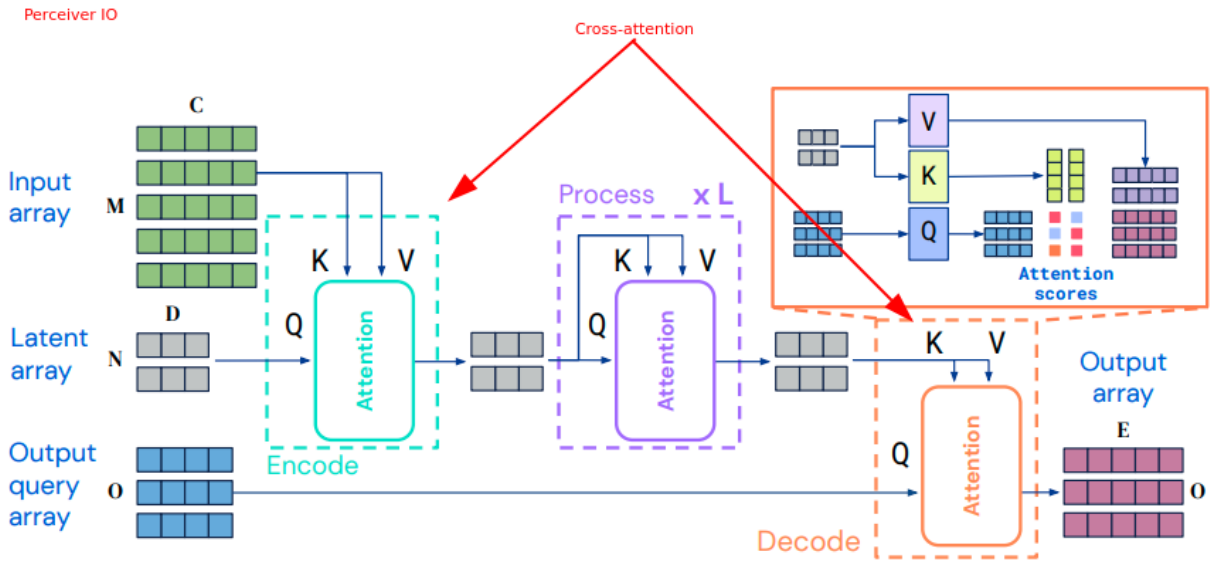


Figure – Application of cross-attention

- **Co-Attention :**
  - **Definition :** Co-attention refers to an attention mechanism that simultaneously considers and aligns information from multiple input sequences or modalities. co-attention can be coarse-grained [12] or fine-grained [13].

12. Coarse-grained attention computes attention on each input, using an embedding of the other input as a query

13. Fine-grained attention evaluates how each element of an input affects each element of the other input

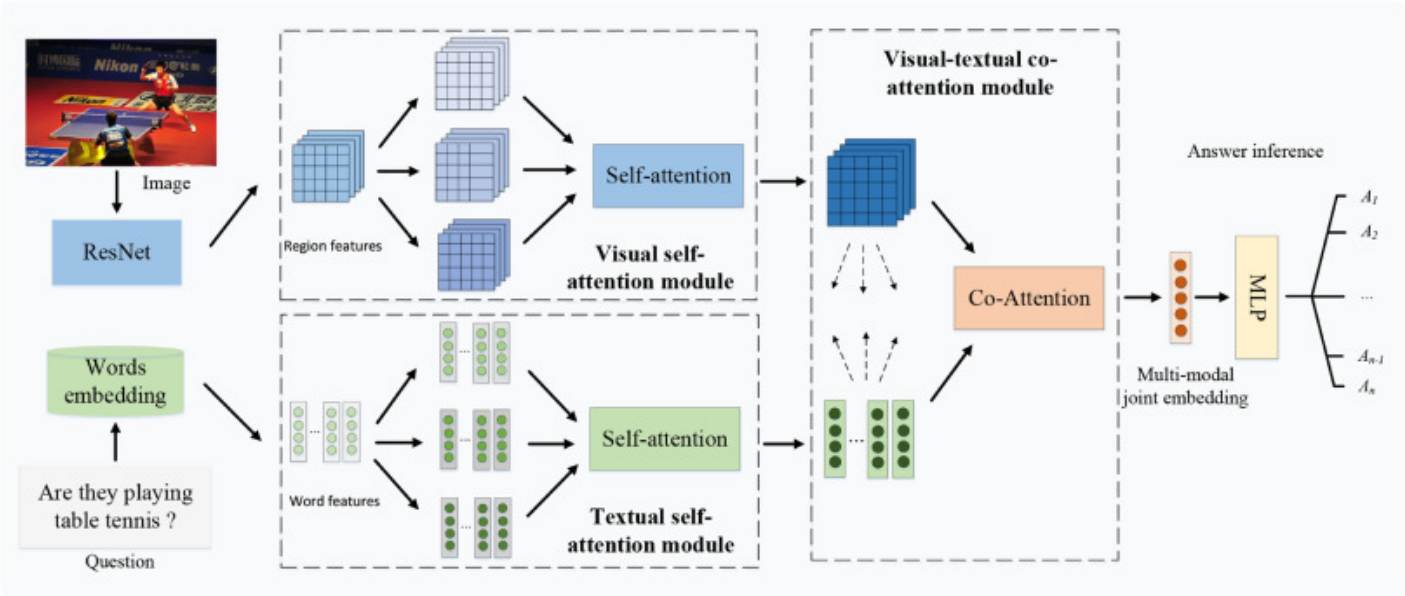


Figure – Application of co-attention

- **Hierarchical Attention :**
  - **Definition :** Hierarchical attention allows the computation of attention weights not only from the original input sequence but also from different abstraction levels (ex. document classification).

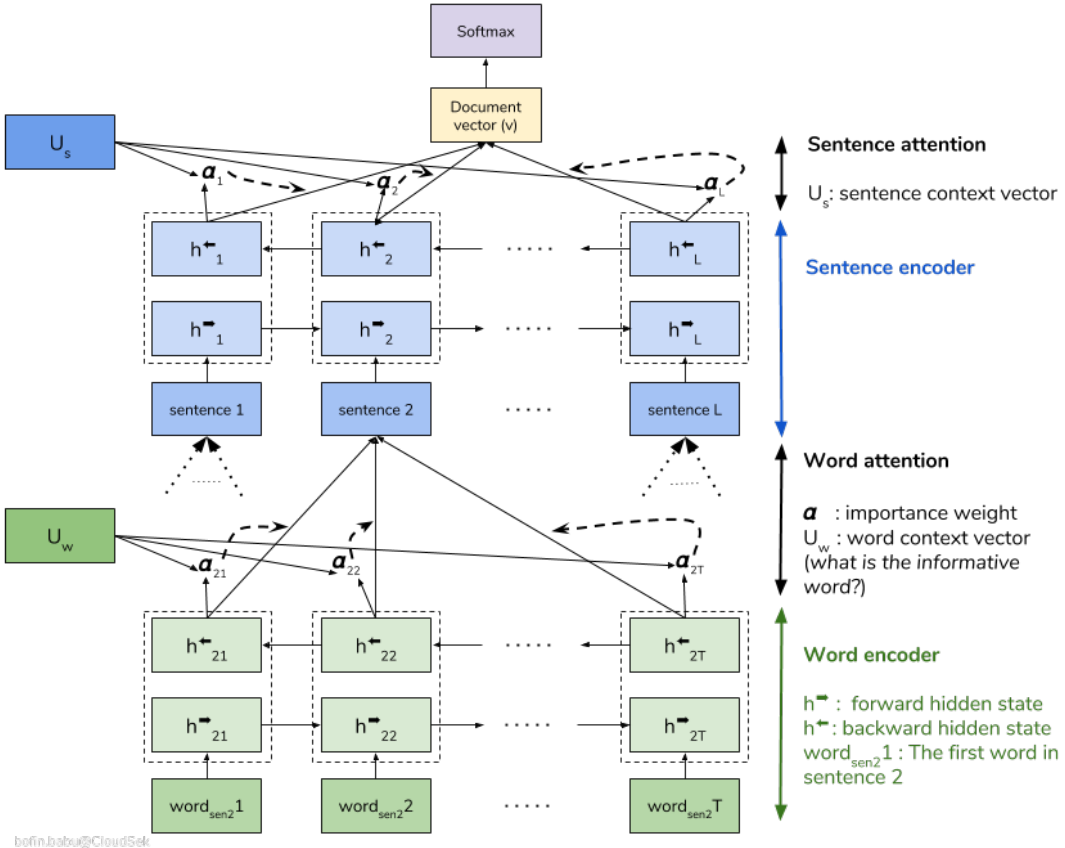


Figure – Application of hierarchical attention

Categorie 4 : Output Representations

- **Single Output Attention :**
  - **Definition :** The energy scores are represented by one and only one vector at each time step.

- **Multi-Head Attention :**
  - **Definition :** An attention mechanism that employs multiple attention heads to capture diverse features and relationships in parallel.

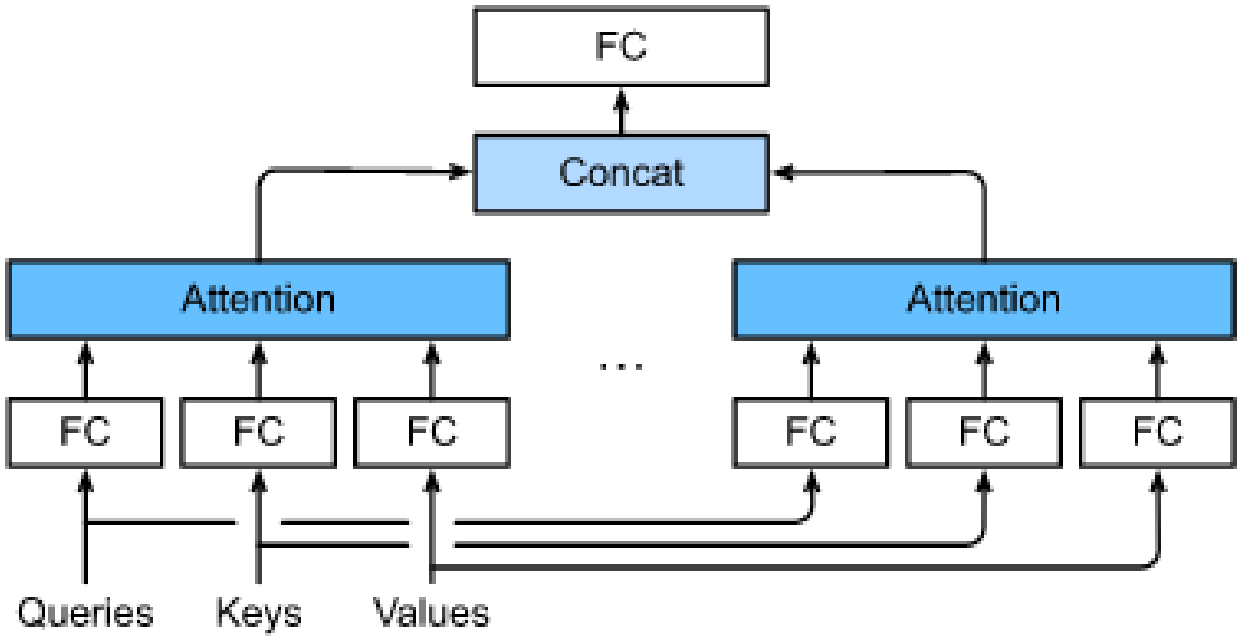


Figure – Application of multi-head attention

- **Multi-Dimensional Attention :**
  - **Definition :** An approach that computes a feature-wise score vector for keys by replacing weight scores vector with a matrix. In this way, the neural network can calculate multiple attention distributions for the same data.

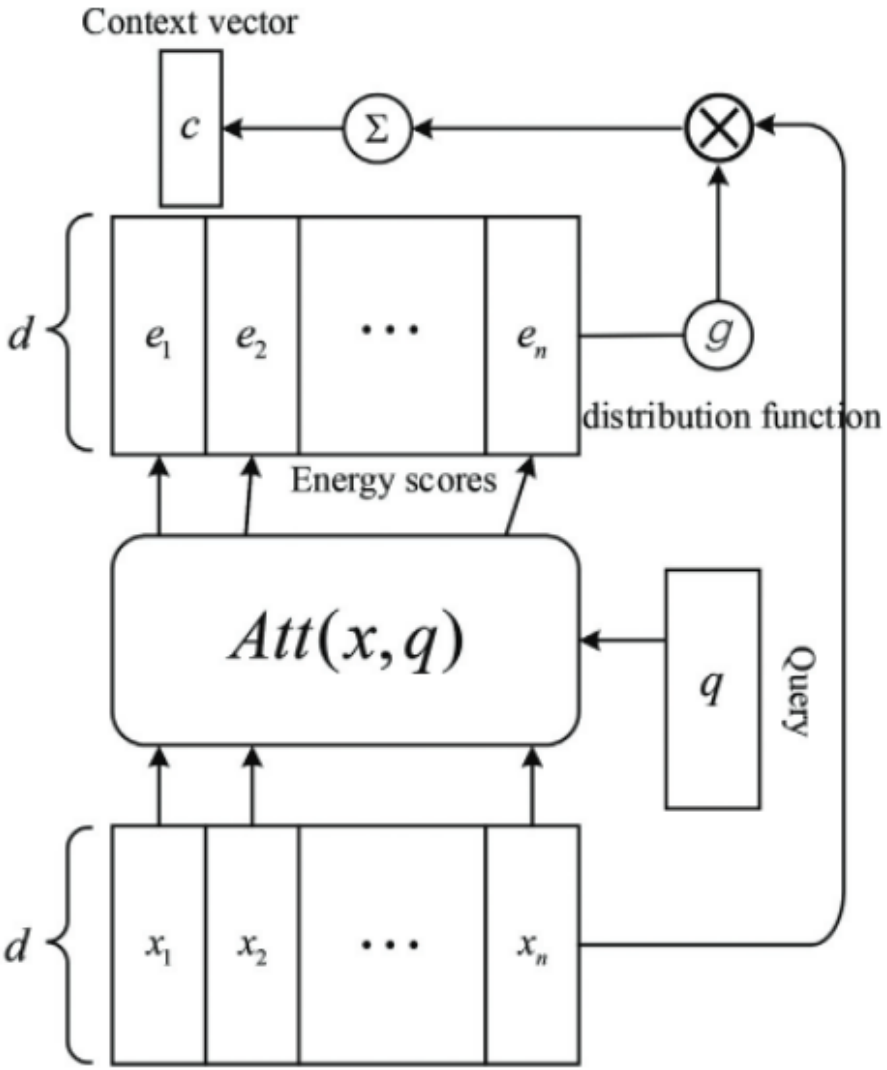
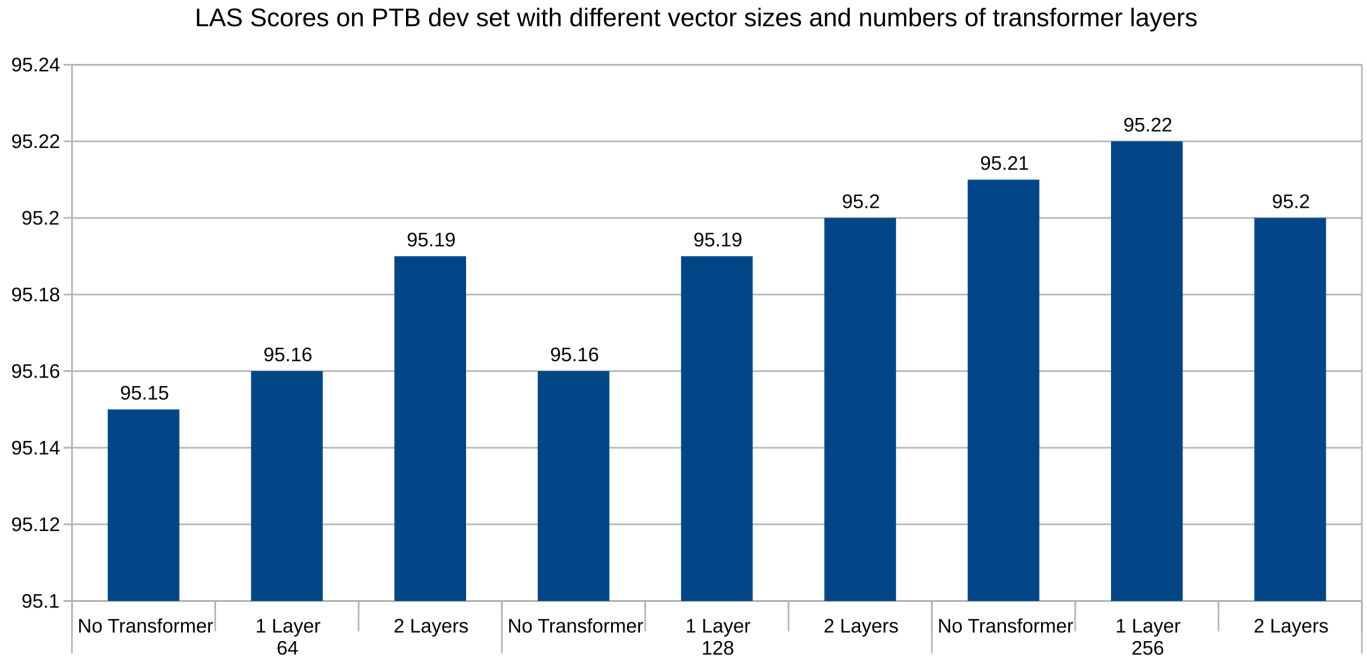
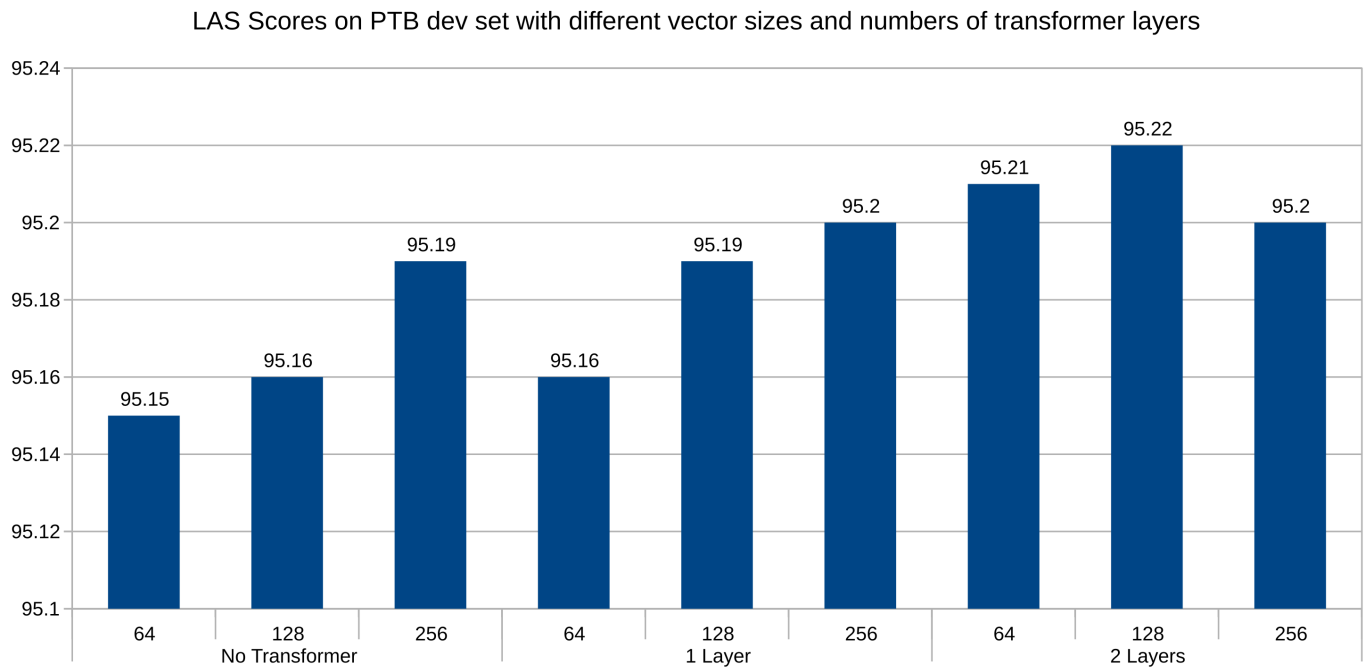


Figure – Application of multi-dimensional attention

Application : Dependency parsing


LAS Scores on PTB dev set with different vector sizes and numbers of transformer layers
95.24
95.22
95.2
95.18
95.16
95.14
95.12
95.1
95.15
95.16
95.19
95.16
95.19
95.2
95.21
95.22
95.2
No Transformer
1 Layer
64
2 Layers
No Transformer
1 Layer
128
2 Layers
No Transformer
1 Layer
256
2 Layers


LAS Scores on PTB dev set with different vector sizes and numbers of transformer layers
95.24
95.22
95.2
95.18
95.16
95.14
95.12
95.1
95.15
95.16
95.19
95.16
95.19
95.2
95.21
95.22
95.2
64
128
256
64
128
256
64
128
256
No Transformer
1 Layer
2 Layers