DIPARTIMENTO DI SCIENZE E TECNOLOGIE

CORSO DI LAUREA IN INFORMATICA APPLICATA
(Machine Learning – Big Data)

Neural Machine Translation

Student :
Ciro Panariello
Matricola 0120000209

Academic Year 2021/2022

Summary

- Introduction to NLP
- What is NMT
- Application I : intro
- Application II : all we need
- Application III : code
- Application IV : Attention Mechanism
- Application V : possible improvements

Introduction to NLP

- **Natural Language Processing**, or **NLP**, is a subfield of Artificial Intelligence research that is focused on developing models and points of interaction between humans and computers based on natural language. This includes text, but also speech-based systems.

- **Why NLP is difficult**

- There is an infinite number of **different ways to arrange words in a sentence**.
- Words can have several meanings and contextual **information is necessary to correctly interpret sentences**. Every language is more or less unique and ambiguous.
- SYNTACTIC & SEMANTIC ANALYSIS : Syntax is the grammatical structure of the text, whereas semantics is the meaning being conveyed. A sentence that is syntactically correct, however, is not always semantically correct.

Introduction to NLP

NLP Applications

- **Chatbox** is used for automatic question answering, designed to understand natural language and deliver an appropriate response through natural language generation.
- **Sentiment analysis** is able to recognize subtle nuances in emotions and opinions – and determine how positive or negative they are.
- **Text Classification** involves automatically understanding, processing, and categorizing unstructured text.
- **Text Extraction** or information extraction, automatically detects specific information in a text, such as names, companies, places, and more. This is also known as named entity recognition.
- **Text Summarization** : it summarizes text, by extracting the most important information.
- **Speech Recognition** is used for transform spoken language into a machine-readable format.
- **Machine Translation** : automatic translation of text from one language to another.

What is NMT

- Machine translation is the task of automatically converting source text in one language to text in another language.

- Machine translation is perhaps one of the most challenging artificial intelligence tasks given the fluidity of human language. Classically, rule-based systems were used for this task, which were replaced in the 1990s with statistical methods. More recently, deep neural network models achieve state-of-the-art results in a field that is aptly named **Neural Machine Translation**.

Application I : intro

- Dataset source : http://www.manythings.org/anki/

- It consists of 300k translated Italian/English sentences.

- Italian vocabulary has 26179 words compared to ~600000 in real world.
- English vocabulary has 12849 words compared to ~1000000 in real world.

- The maximum length of an Italian sentence is 12 words.
- The maximum length of an English sentence is 13 words.

Application II : all we need

Prerequisites

- Recurrent Neural Network
- LSTM

In NLP we use

- Word Embedding
- Encoder/Decoder model

Word Embedding

- One **hot encoding** is a simple representation for the input data.

Example :

- The dataset consists of 8 words and we assign a unique code to each word.

I	ate	an	apple	and	played	the	piano
1	2	3	4	5	6	7	8

- The word “I” is at position 1, so its one-hot vector representation would be [1, 0, 0, 0, 0, 0, 0, 0]. Similarly, the word “ate” is at position 2, so its one-hot vector would be [0, 1, 0, 0, 0, 0, 0, 0].

Word Embedding

- The one-hot embedding matrix for the example text would look like this:

	1	2	3	4	5	6	7	8
I	1	0	0	0	0	0	0	0
ate	0	1	0	0	0	0	0	0
an	0	0	1	0	0	0	0	0
apple	0	0	0	1	0	0	0	0
and	0	0	0	0	1	0	0	0
played	0	0	0	0	0	1	0	0
the	0	0	0	0	0	0	1	0
piano	0	0	0	0	0	0	0	1

We have 2 major problems :

1. each pattern will have a number of features equal to the number of words in the dictionary → not computationally efficient.
2. we lose the semantic relationship that exists between the words of the sentence.

Word Embedding

- The purpose of word embedding is that **semantic similar words have a shorter distance** (euclidean, cosine, etc.) than words that have no semantic relationship. For example, words like "mom" and "dad" should be closer than the words "mom" and "ketchup" or "dad" and "butter".

- The goal is to have a **word vector** for each word of the dataset in order to properly identify its relationship with all the other words.

- We have several neural network based approaches to train word embedding, including CBOW and Skip-Gram models.

Encoder/Decoder

- The **encoder/decoder model** is a way to use recurrent neural networks for sequence-sequence prediction problems.

- The approach involves two recurrent neural networks, one to **encode the input sequence**, called the encoder, and a second to **decode the input sequence encoded** in the output sequence, called the decoder.

- The encoder/decoder architecture is widely used in NMT → in general, widely used when dealing with sequential data in which the input and output can have a different sequence length; it is precisely the case of the NMT where we have that the input (sentence to be translated) will most likely have a different length from the output (translated sentence).

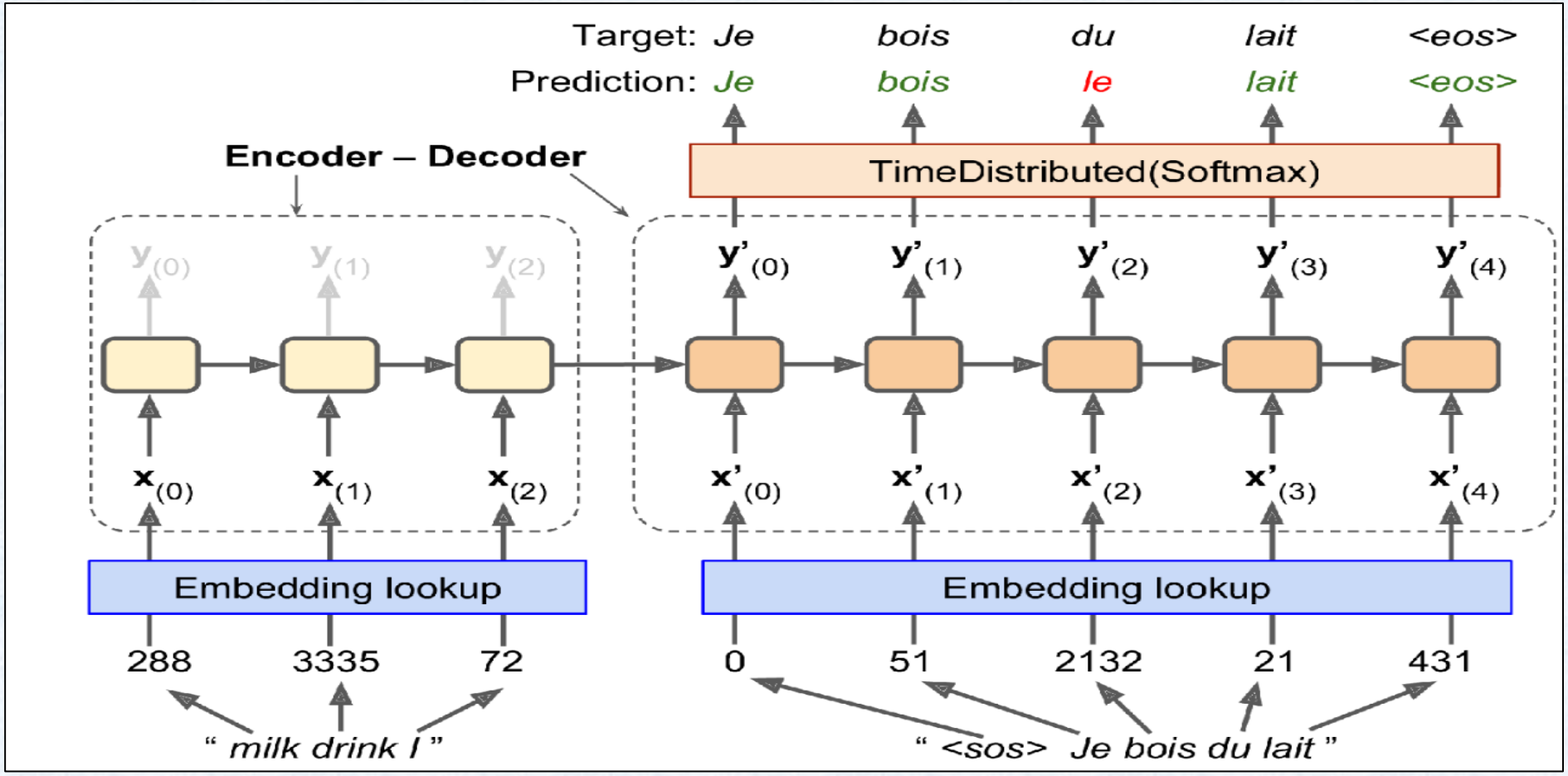
Encoder/Decoder

There are two ways to implement an encoder/decoder model :

1. Using repeat encoded vector
2. Using Teacher forcing

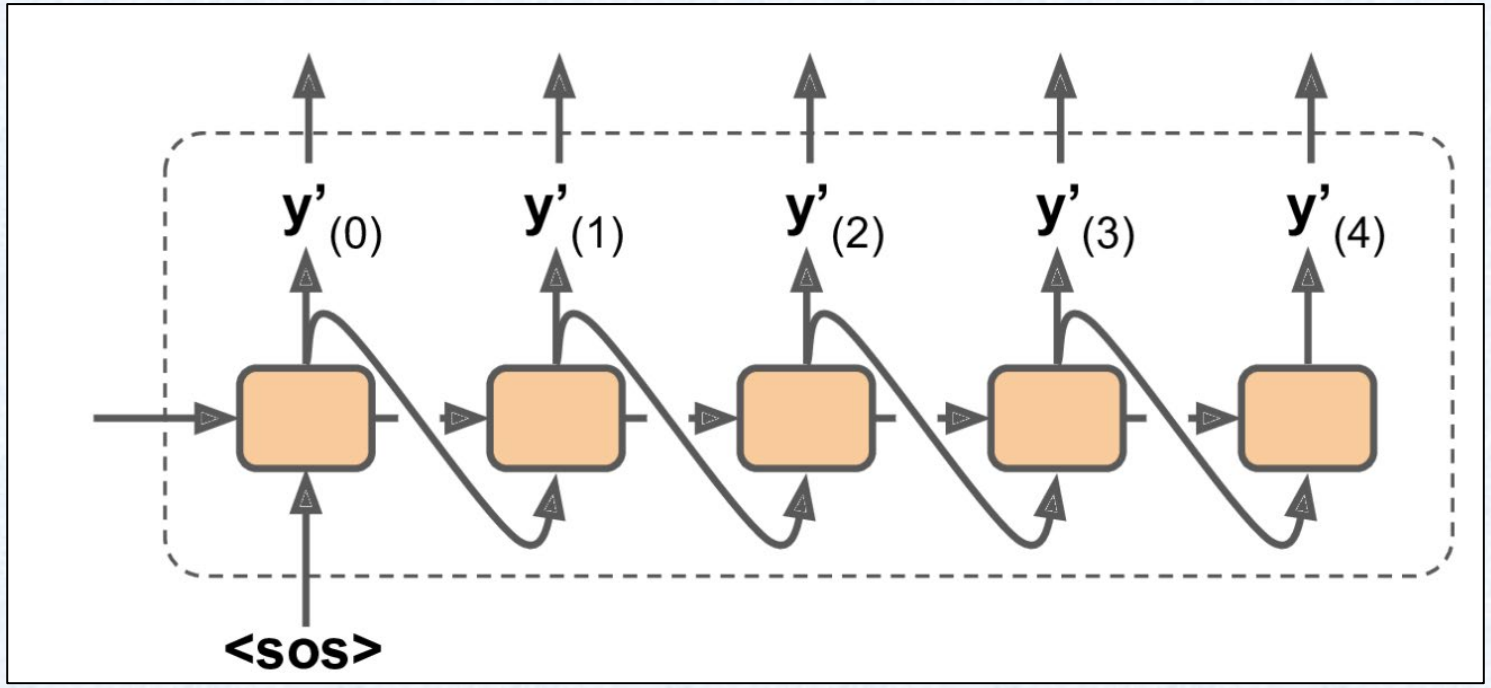
Encoder/Decoder

- Using **Teacher forcing** → Training phase



Encoder/Decoder

- Using **Teacher forcing** → Inference phase



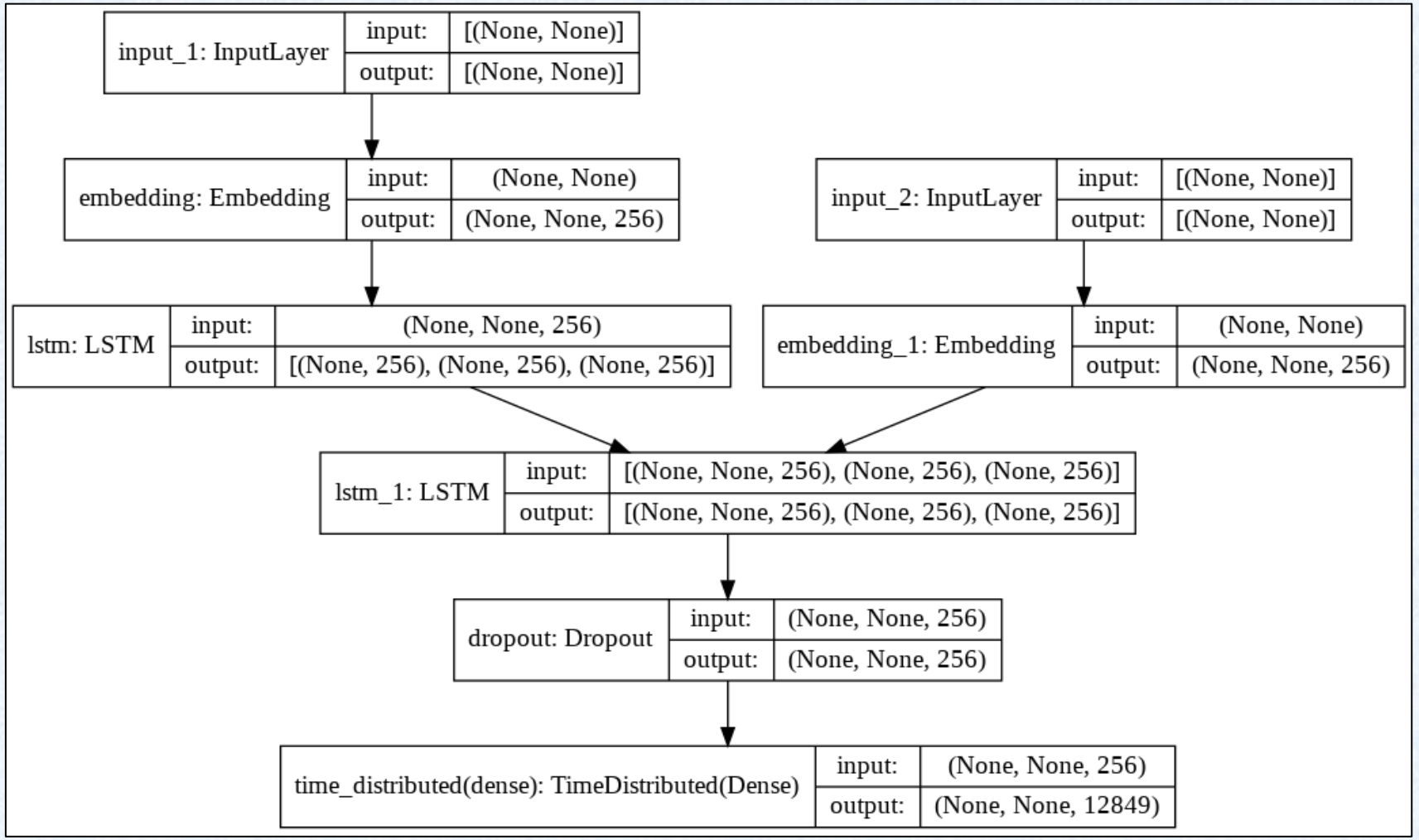
Application III : code

- https://github.com/Dantekk/Neural-Machine-Translation-with-attention-mechanism

- There are two versions of the project :

1. OOP
2. Jupyter notebook : https://github.com/Dantekk/Neural-Machine-Translation-with-attention-mechanism/blob/main/Neural_Machine_Translation.ipynb

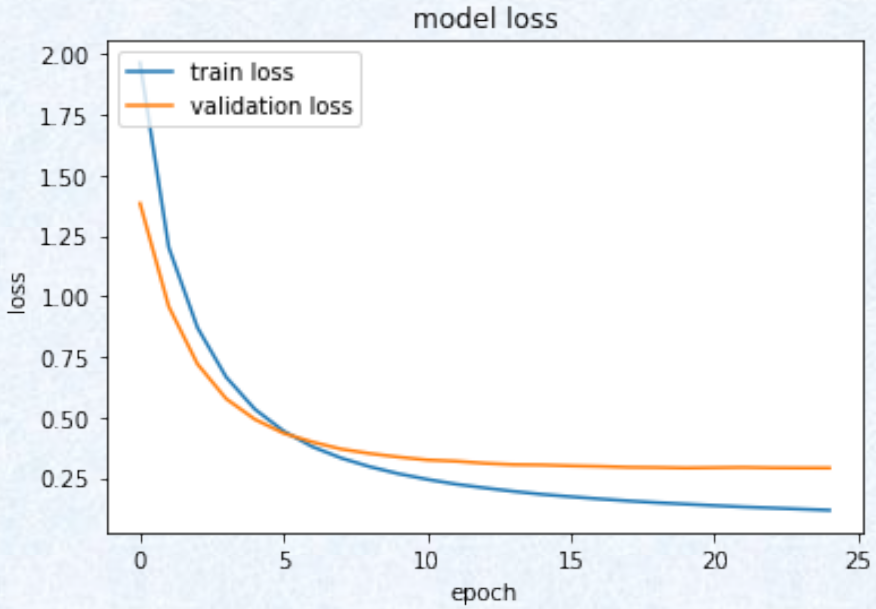
Application III : code

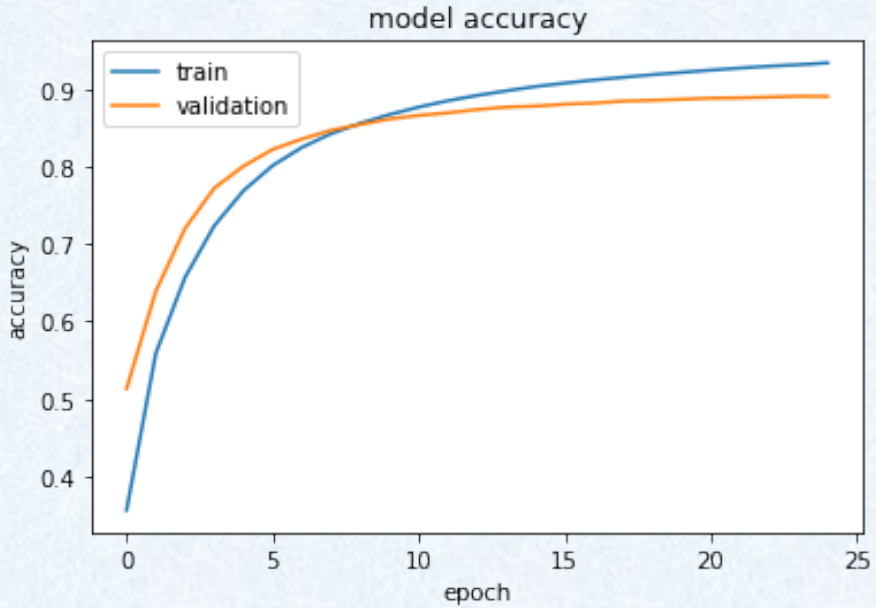


Application III : code

- **Loss :** sparse_categorical_crossentropy
- **Optimizer :** Adam
- **Batchsize :** 128

[epoch 21] : loss: 0.1336 - accuracy: 0.9249 - val_loss: 0.2882 - **val_accuracy: 0.8887**





Application III : code

```
Input  sentence :  come stai?
Output sentence :  how are you ? <end>

Input  sentence :  quanti anni hai?
Output sentence :  how old are you ? <end>

Input  sentence :  come ti chiami?
Output sentence :  whats your name ? <end>

Input  sentence :  è una bellissima giornata
Output sentence :  its a beautiful day <end>

Input  sentence :  sei una ragazza pericolosa
Output sentence :  youre a dangerous girl <end>

Input  sentence :  ho studiato duramente per superare l'esame
Output sentence :  i studied hard in order to pass the examination <end>

Input  sentence :  mia madre dice sempre che sono bello
Output sentence :  my mother always says hes handsome <end>
```

Application III : code

```
Input  sentence :  conosco una ragazza che parla molto bene il francese
Target sentence :  i know a girl who speaks french very well
Output sentence :  i know a girl who speaks french very well <end>

Input  sentence :  che cosa fareste se questo fosse il vostro ultimo giorno?
Target sentence :  what would you do if this was your last day ?
Output sentence :  what would you do if this was your last day ? <end>

Input  sentence :  le previsioni del tempo dicono che farà bello domani
Target sentence :  the weather forecast says it will be fine tomorrow
Output sentence :  the weather say you plan now think well <end>

Input  sentence :  dove posso comprare una chitarra da queste parti ?
Target sentence :  where can i buy a guitar around here?
Output sentence :  where can i buy a nearby area ? <end>
```

Application IV : Attention Mechanism

- The **attention mechanism** has changed the way we work with deep learning algorithms.
- Fields such as NLP and even Computer Vision have been revolutionized by the attention mechanism.
- It spawned the rise of so many recent breakthroughs in NLP, including the Transformer architecture and Google's BERT.

Application IV : Attention Mechanism

What is Attenion?

- In psychology, attention **is the cognitive process of selectively focusing on one or a few things while ignoring others**.
- You are looking at a group photo of your first school. Typically, there will be a group of children. Now, if someone asks the question, “How many people are there?”
- How will you answer?

Application IV : Attention Mechanism

What is Attenion?

- Just by counting heads, right? There is no need to consider other things in the photo.
- This is precisely the mechanism of attention → concentrate on the parts that we consider important and exclude everything else.

Application IV : Attention Mechanism

Encoder/Decoder model problems :

1. The LSTM output of the last step of the encoder is used to bind the encoder to the decoder → this is called the **context vector** and we can consider it as a summary of the encoder. All the other output steps of the encoder are ignored → we do not exploit them in any way in the decoder.
2. If the encoder makes a wrong summary, the translation will also be wrong. And in fact it has been observed that the encoder creates a bad summary when trying to understand longer sentences. It is called the **long-range dependence problem** of RNN/LSTM.
3. There is no way to give some input words more importance than others when translating the sentence.

Basically, starting with a context vector we are asking to get too much to the decoder.

Application IV : Attention Mechanism

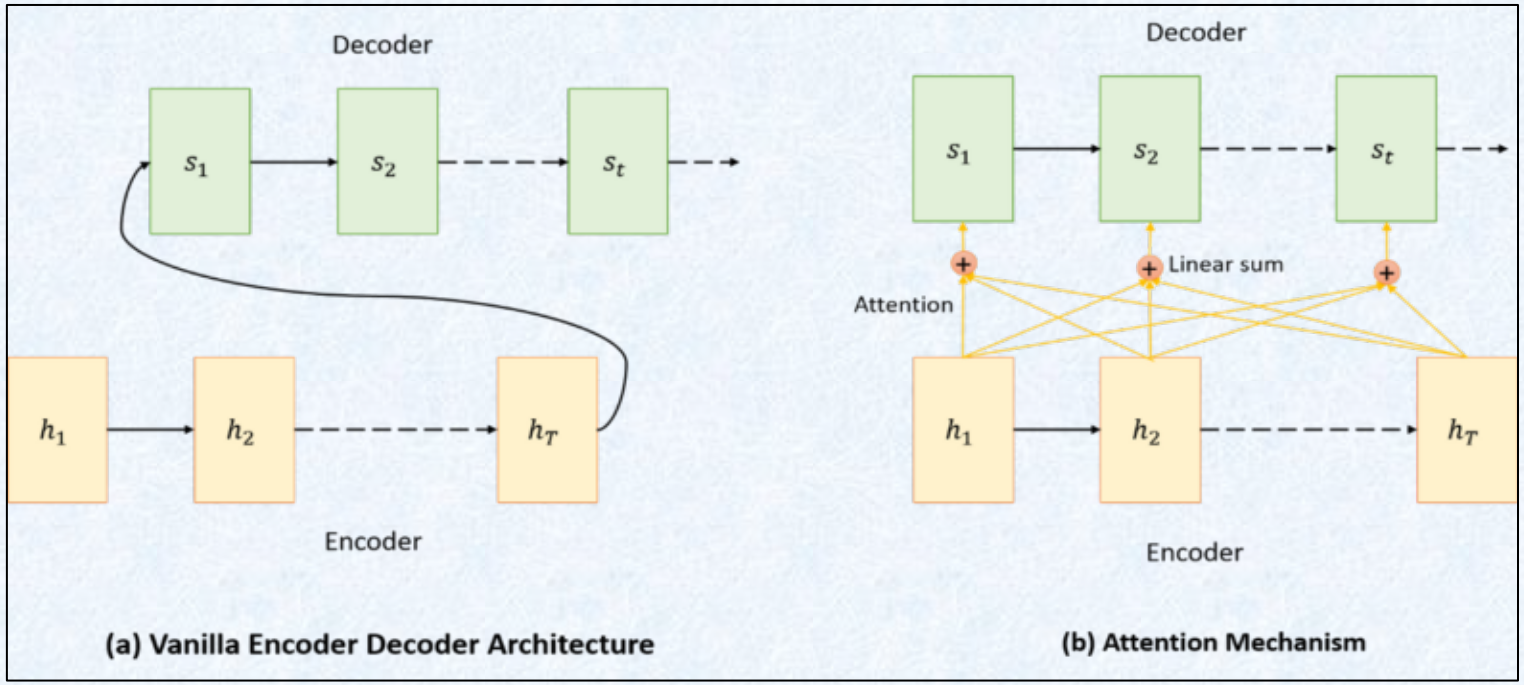
- The **attention mechanism** was introduced in 2015 in this article by **Bahdanau** (KyungHyun Cho, Yoshua Bengio).

- https://arxiv.org/pdf/1409.0473.pdf

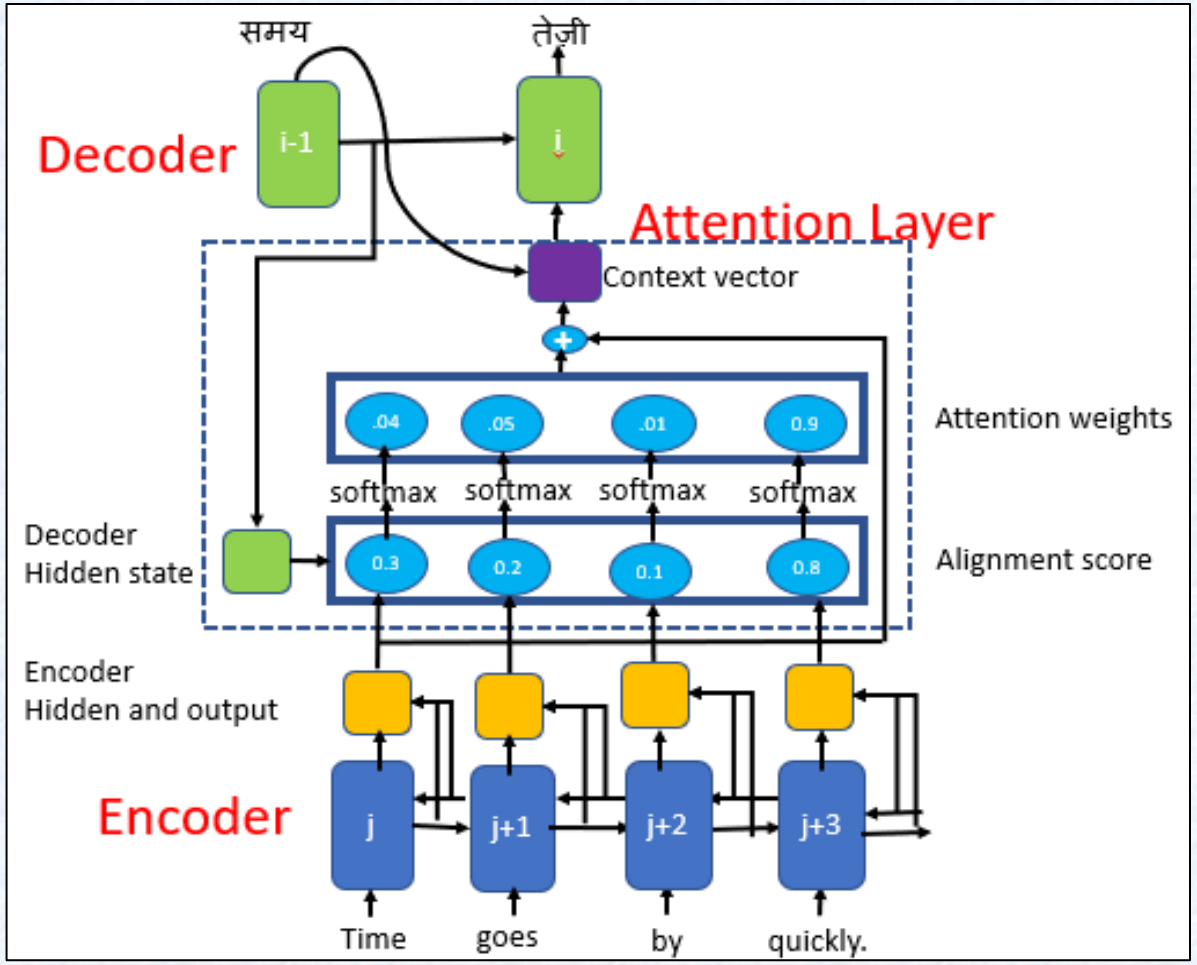
- It was the first time that the concept of attention for neural networks was approached, for this reason **the author used it in the paper with great caution**.

Application IV : Attention Mechanism

With the attention mechanism, **the decoder at each step uses all the hidden states of the entire encoder sequence** (and not just the last step of the encoder) to make predictions, unlike the vanilla encoder/decoder approach.



Application IV : Attention Mechanism



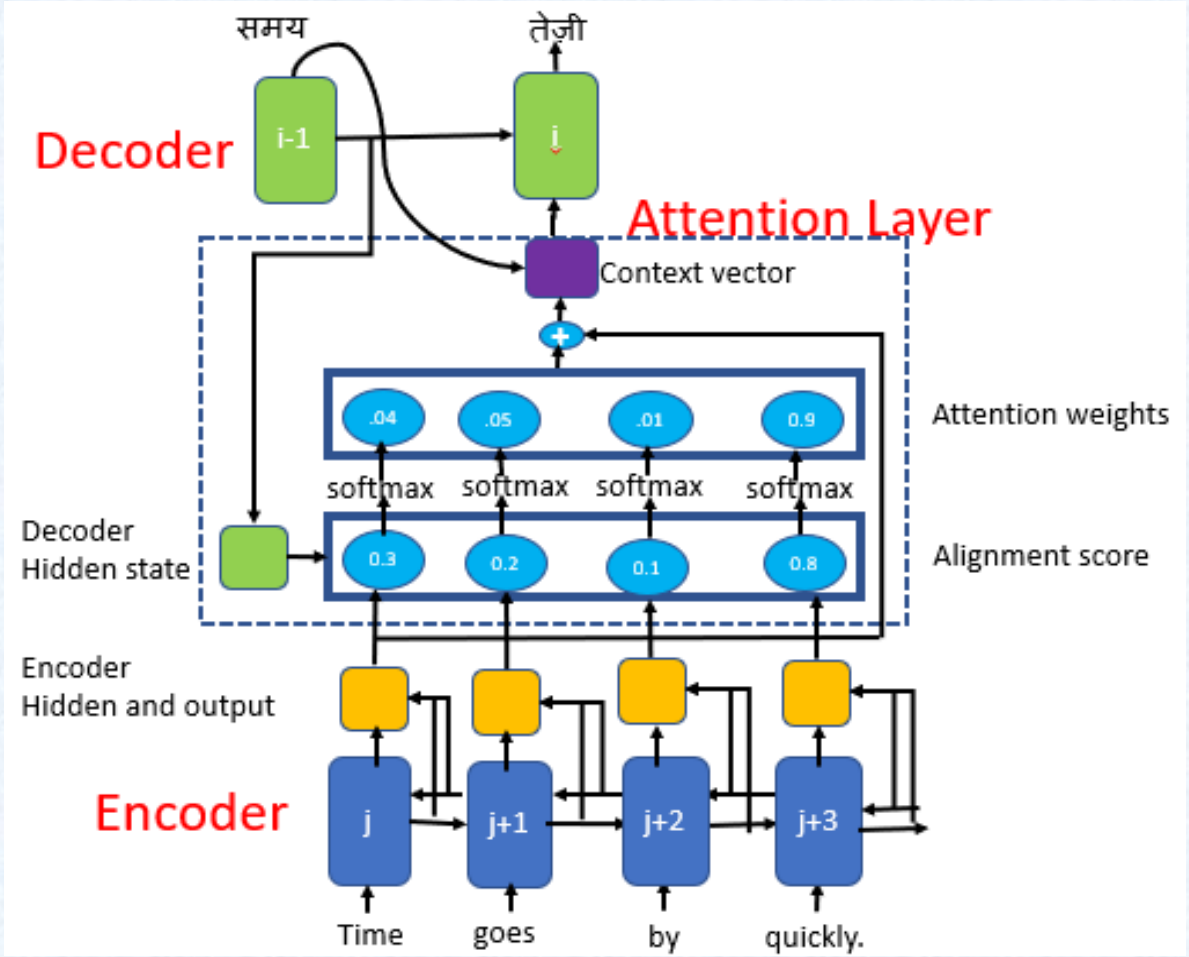
A neural network is used to calculate the vector of the **alignment scores** :

$$e_{ij.} = a(s_{i-1}, h_j) \qquad \text{Alignment Score}$$

Let's apply a softmax function to the alignment scores to get **attention weights**:

$$\alpha_{i_j} = \exp(e_{ij}) \Big/ \sum_{k=1}^{Tx} \exp(e_{ik}) \qquad \text{Attention weight}$$

Application IV : Attention Mechanism



- **alignment scores** :

$$e_{ij.} = a(s_{i-1}, h_j) \qquad \text{Alignment Score}$$

- **attention weights**:

$$\alpha_{i_j} = \exp(e_{ij}) \Big/ \sum_{k=1}^{Tx} \exp(e_{ik}) \qquad \text{Attention weight}$$

Finally, we calculate the **context vector** as a linear combination of the outputs of each step of the encoder for the attention weights just calculated.

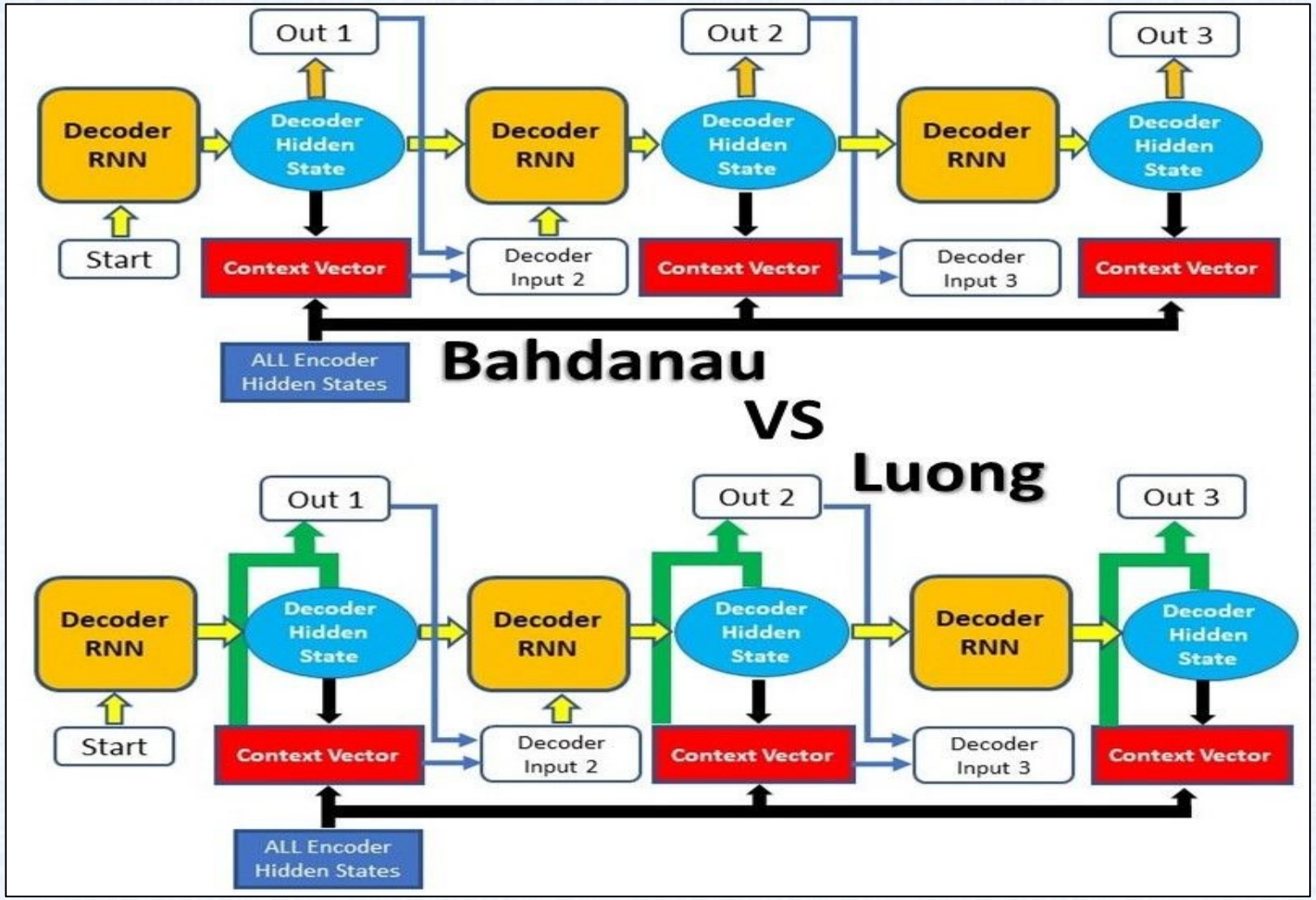
$$C_i = \sum_{j=1}^{Tx} \alpha_{ij}\, h_j \qquad \text{Context vector}$$

Application IV : Attention Mechanism

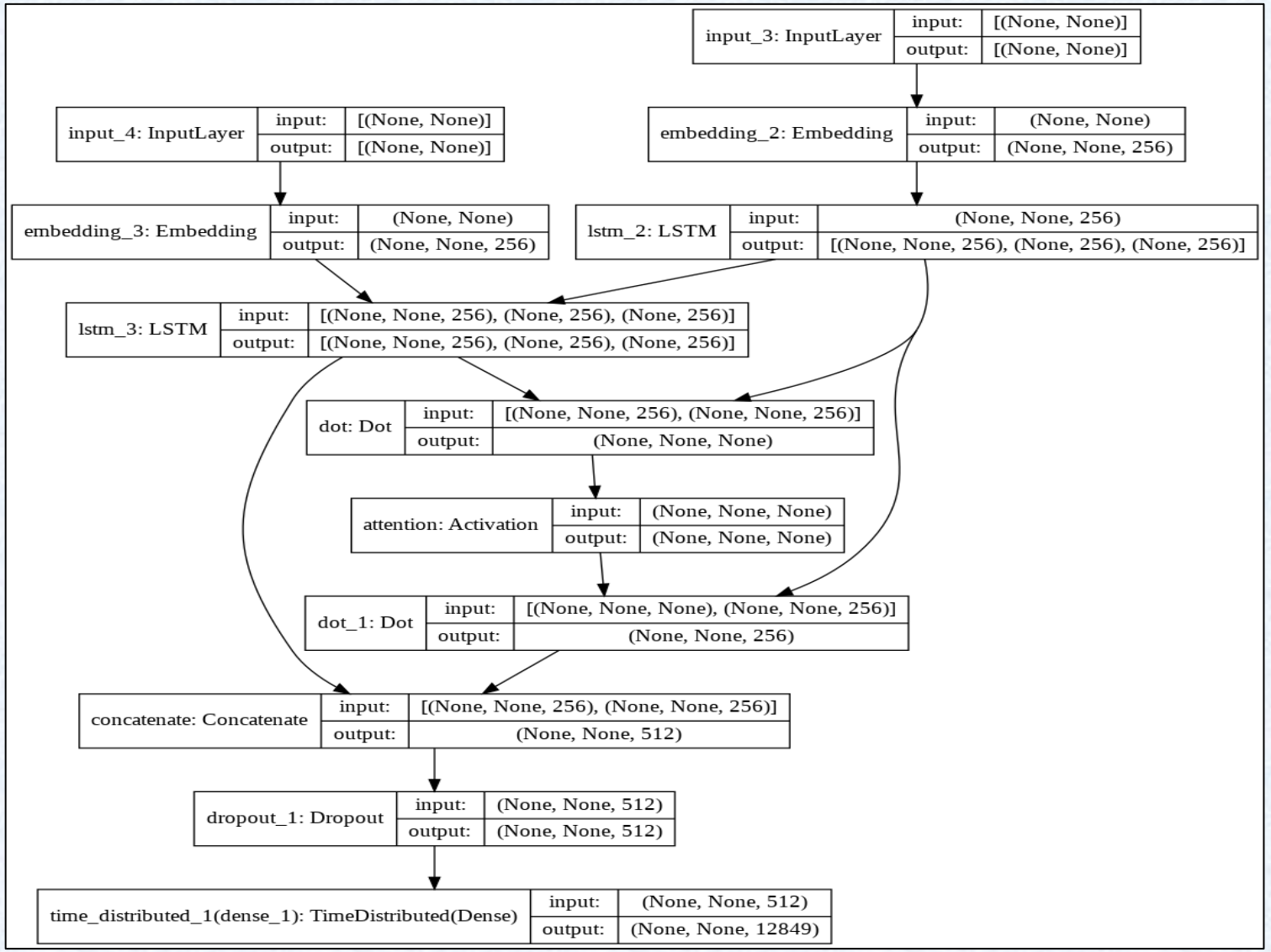
- **Bahdanau attention** has a high overhead → comparable results but with lower computational cost are obtained with **Luong Attention**.

- https://arxiv.org/pdf/1508.04025.pdf

Application IV : Attention Mechanism



Application IV : Attention Mechanism



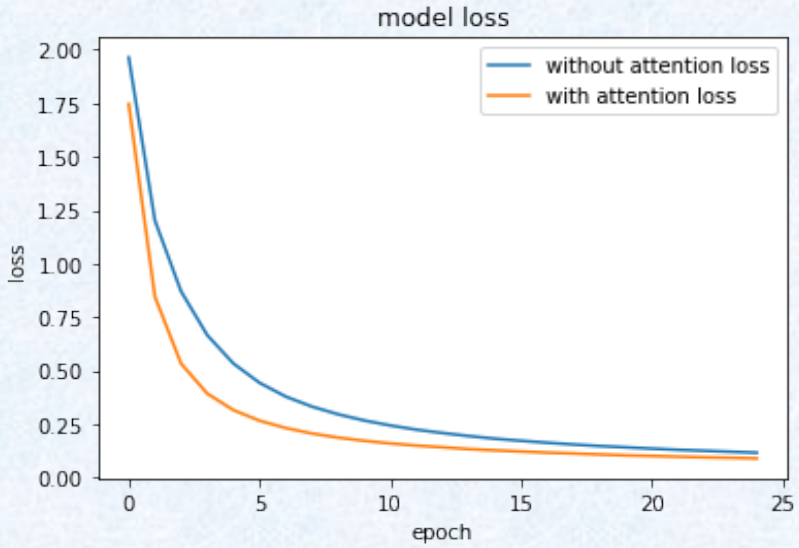
Application IV : Attention Mechanism

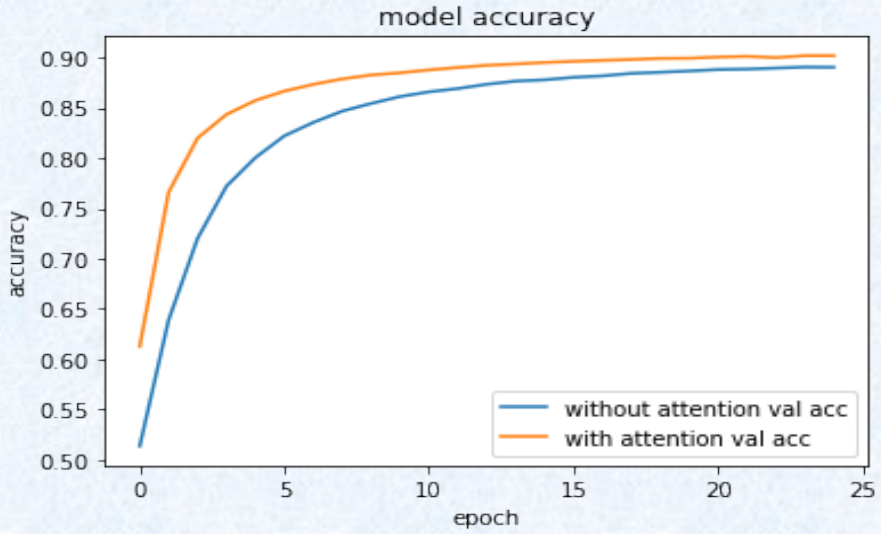
- **Loss :** sparse_categorical_crossentropy
- **Optimizer :** Adam
- **Batchsize :** 128

The best result was :

```
Without attention : [epoch 21] : loss: 0.1336 - accuracy: 0.9249 - val_loss: 0.2882 - val_accuracy: 0.8887

With attention : [epoch 16] : loss: 0.1213 - accuracy: 0.9323 - val_loss: 0.2697 - val_accuracy: 0.9162
```





Application IV : Attention Mechanism

```
Input sentence              : mia madre dice sempre che sono bello
Target sentence             : my mother always says im handsome
Without attention sentence : my mother always says hes handsome
With attention sentence     : my mother always says im handsome <end>

Input sentence              : le previsioni del tempo dicono che farà bello domani
Target sentence             : the weather forecast says it will be fine tomorrow
Without attention sentence : the weather say you plan now think well
With attention sentence     : the weather report says its going to good tomorrow <end>

Input sentence              : dove posso comprare una chitarra da queste parti ?
Target sentence             : where can i buy a guitar around here ?
Without attention sentence : where can i buy a nearby area ?
With attention sentence     : where can i buy a guitar around here ? <end>
```

Application V : possible improvements

- We have faced the NMT task using an encoder/decoder architecture with teacher forcing and we have seen the improvements that can be obtained by introducing the attention mechanism.

Possible improvements

- ✓ Improve the dataset.
- ✓ Parameter tuning.
- ✓ Use the latest attention mechanism → Transformers.
- ✓ **Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks :** https://arxiv.org/pdf/1506.03099.pdf (2015, Bengio at al)
  GOAL? Decrease discrepancy between training and inference.

Thanks for your attention