



Natural Language Processing with Deep Learning

BÀI 5: MACHINE TRANSLATION



https://csc.edu.vn/data-science-machine-learning/natural-language-processing-with-deep-learning_293



MACHINE TRANSLATION



I. Tổng quan Machine Translation

II. Mô hình Seq2Seq

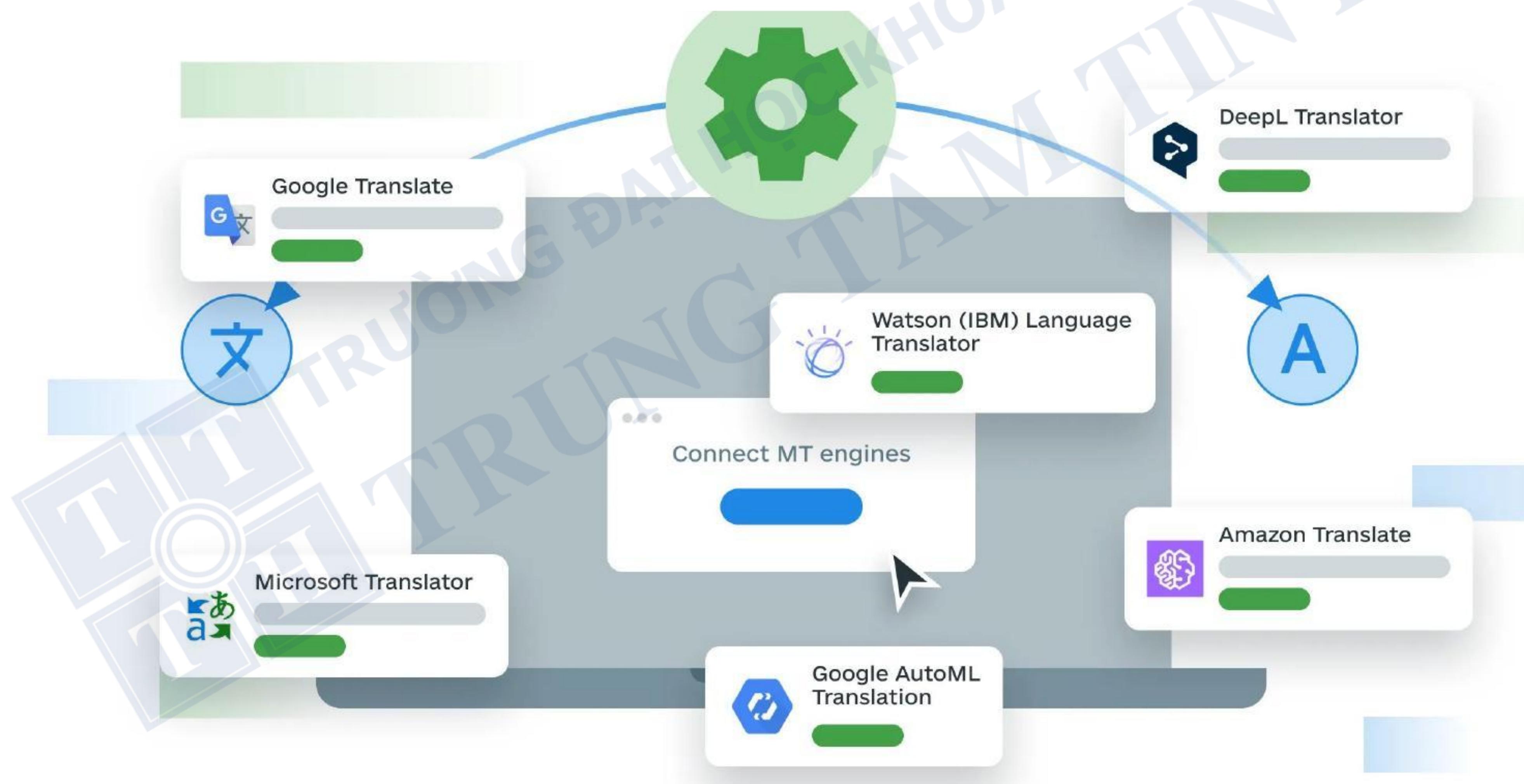
III. Mô hình Seq2Seq + Attention

IV. Neural Machine Translation

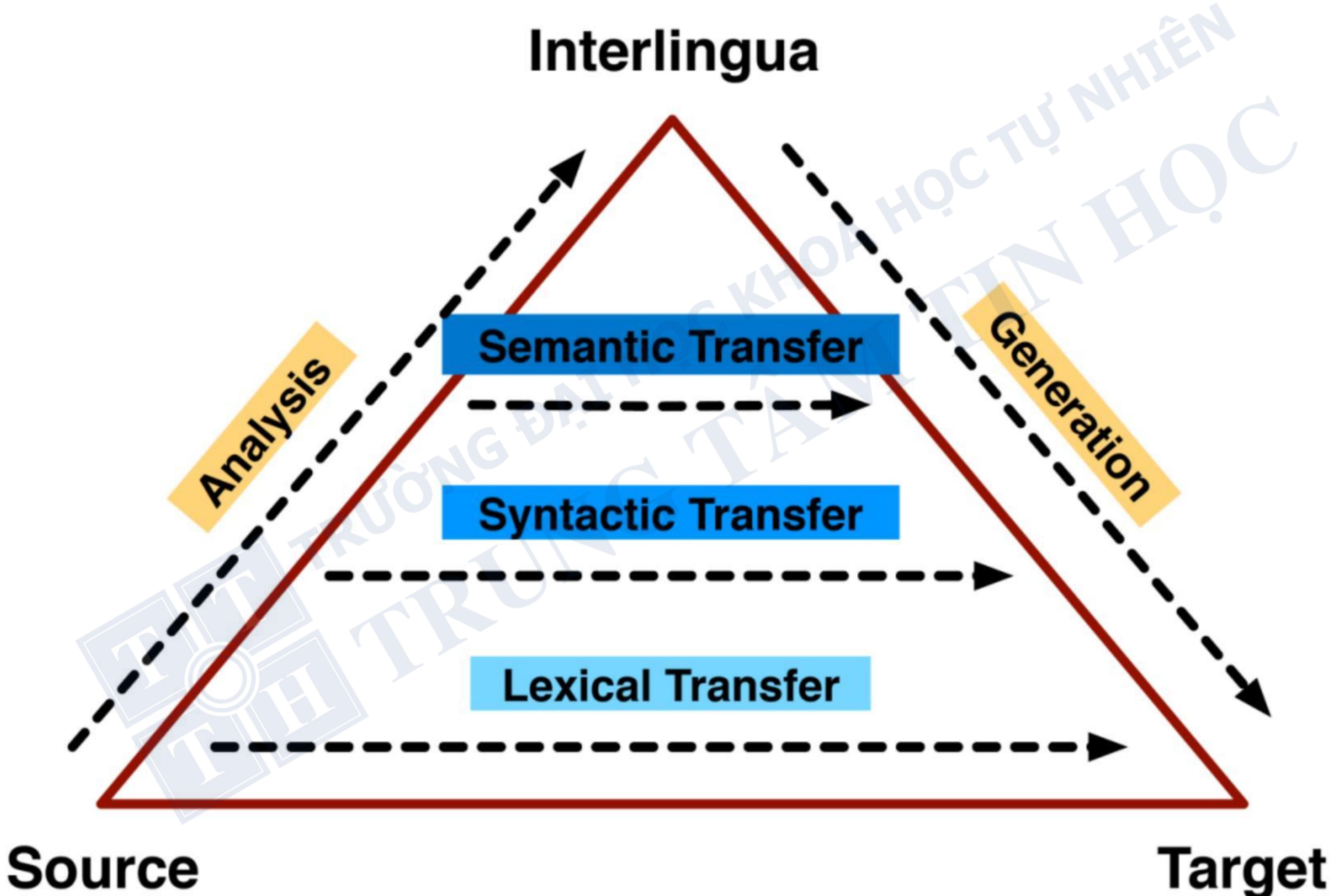


Tổng quan Machine Translation

Machine translation là quá trình tự động dịch đoạn văn bản từ một ngôn ngữ gốc (*source language*) sang một ngôn ngữ đích (*target language*).



Tổng quan Machine Translation





Tổng quan Machine Translation

Làm thế nào để chuyển đổi các từ?

Để chuyển các từ từ ngôn ngữ gốc (source) sang ngôn ngữ đích (target) → Dùng ma trận chuyển đổi **Transformation Matrix**.

Transformation matrix được xác định bằng CT sau:

$$\operatorname{argmax}_y P(x|y)P(y)$$

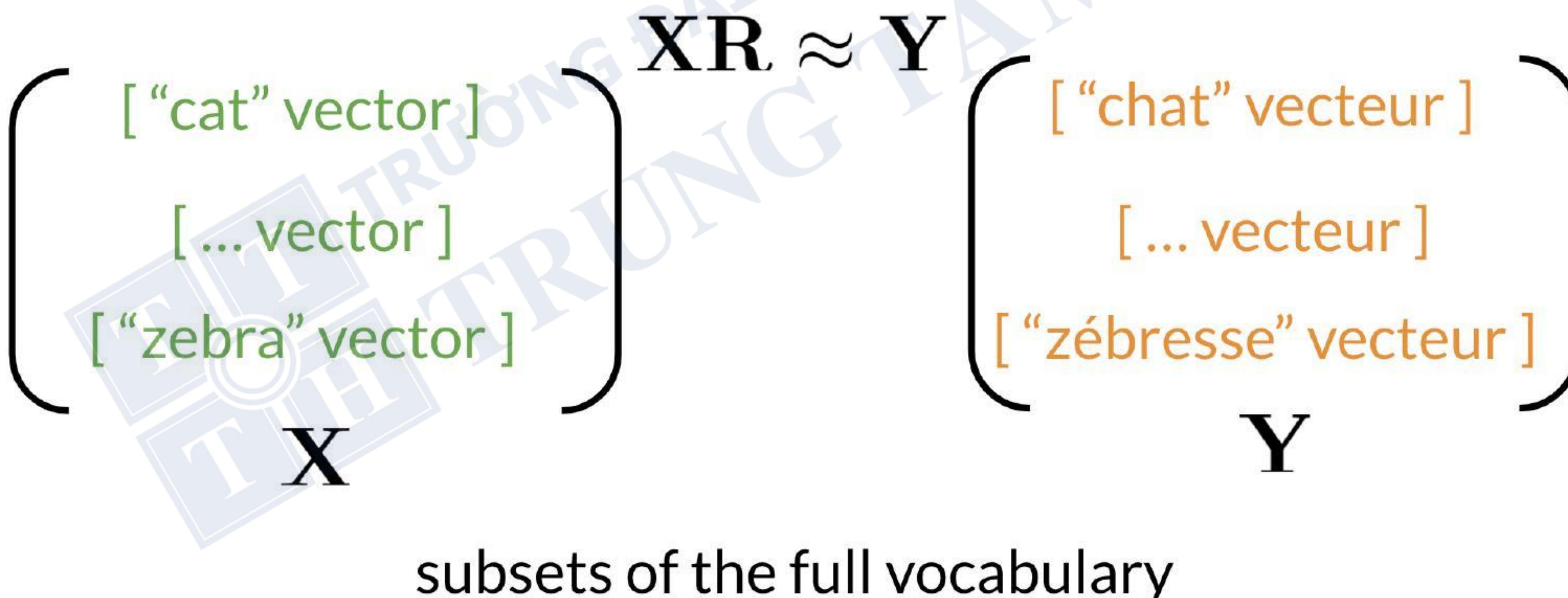
Translation model Language model



Tổng quan Machine Translation

Transformation Matrix

- **X** là ma trận word vector của các từ tiếng Anh.
- **Y** là ma trận word vector của các từ tiếng Pháp.
- **R** là ma trận mapping/ transformation.

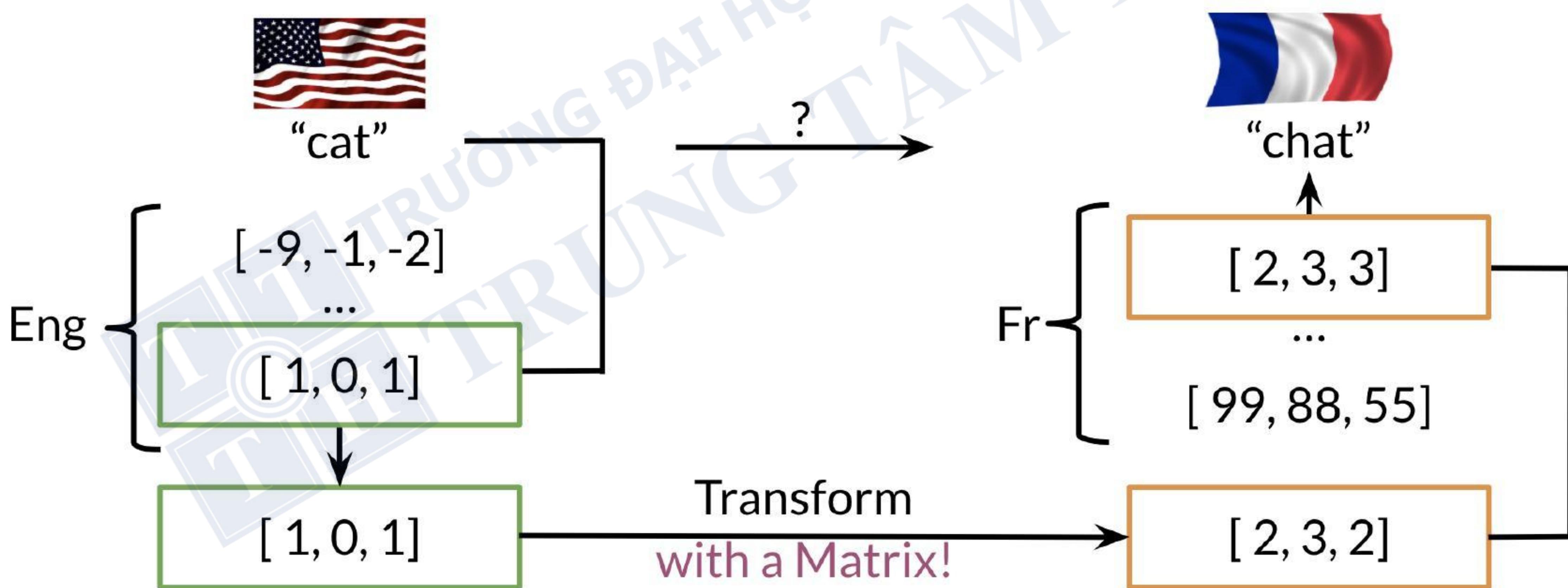




Tổng quan Machine Translation

Transformation Matrix

- **X** là ma trận word vector của các từ tiếng Anh.
- **Y** là ma trận word vector của các từ tiếng Pháp.
- **R** là ma trận mapping/ transformation.



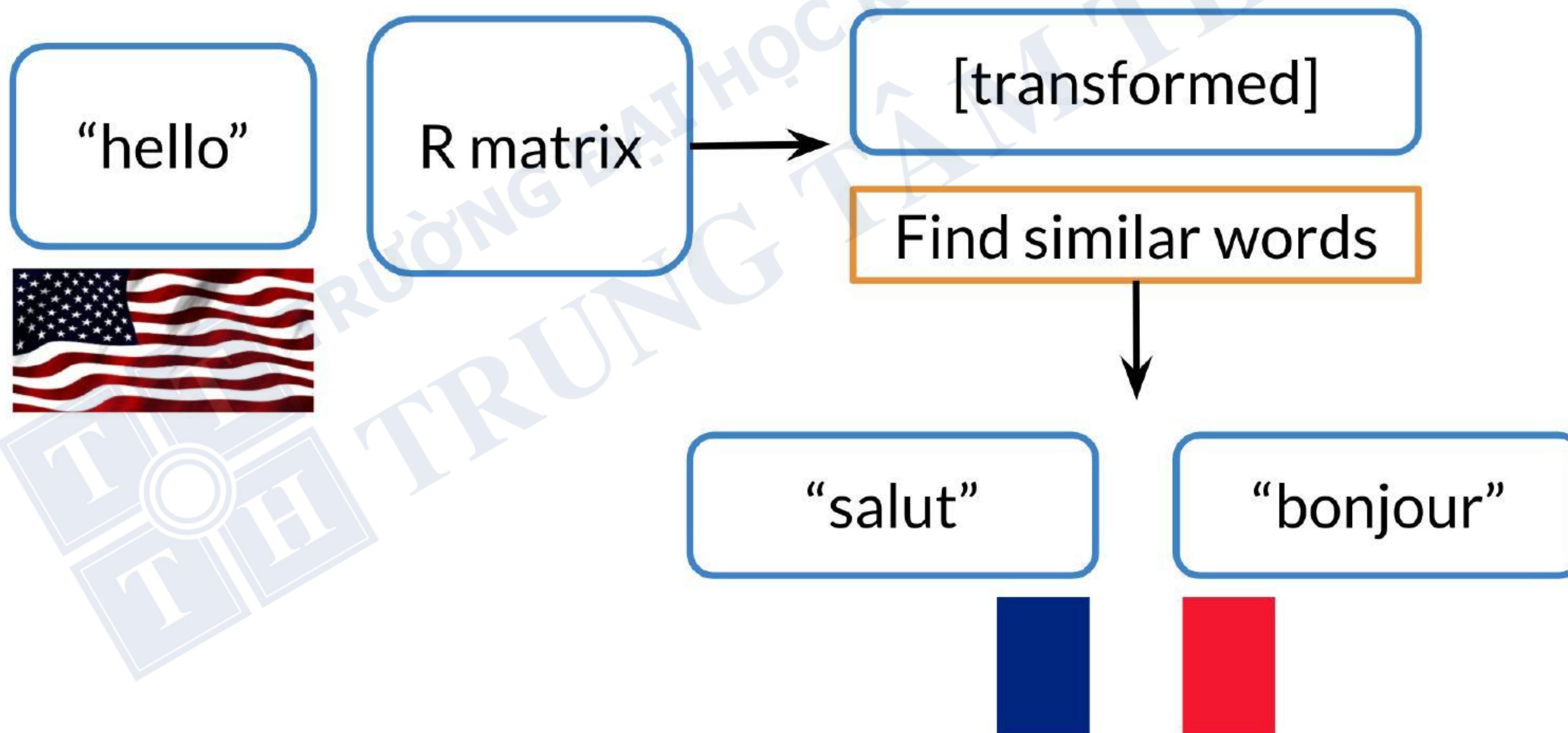


Tổng quan Machine Translation

K-nearest Neighbors

Sau khi có output của X.R

→ Tìm word vector **tương đồng nhất**.





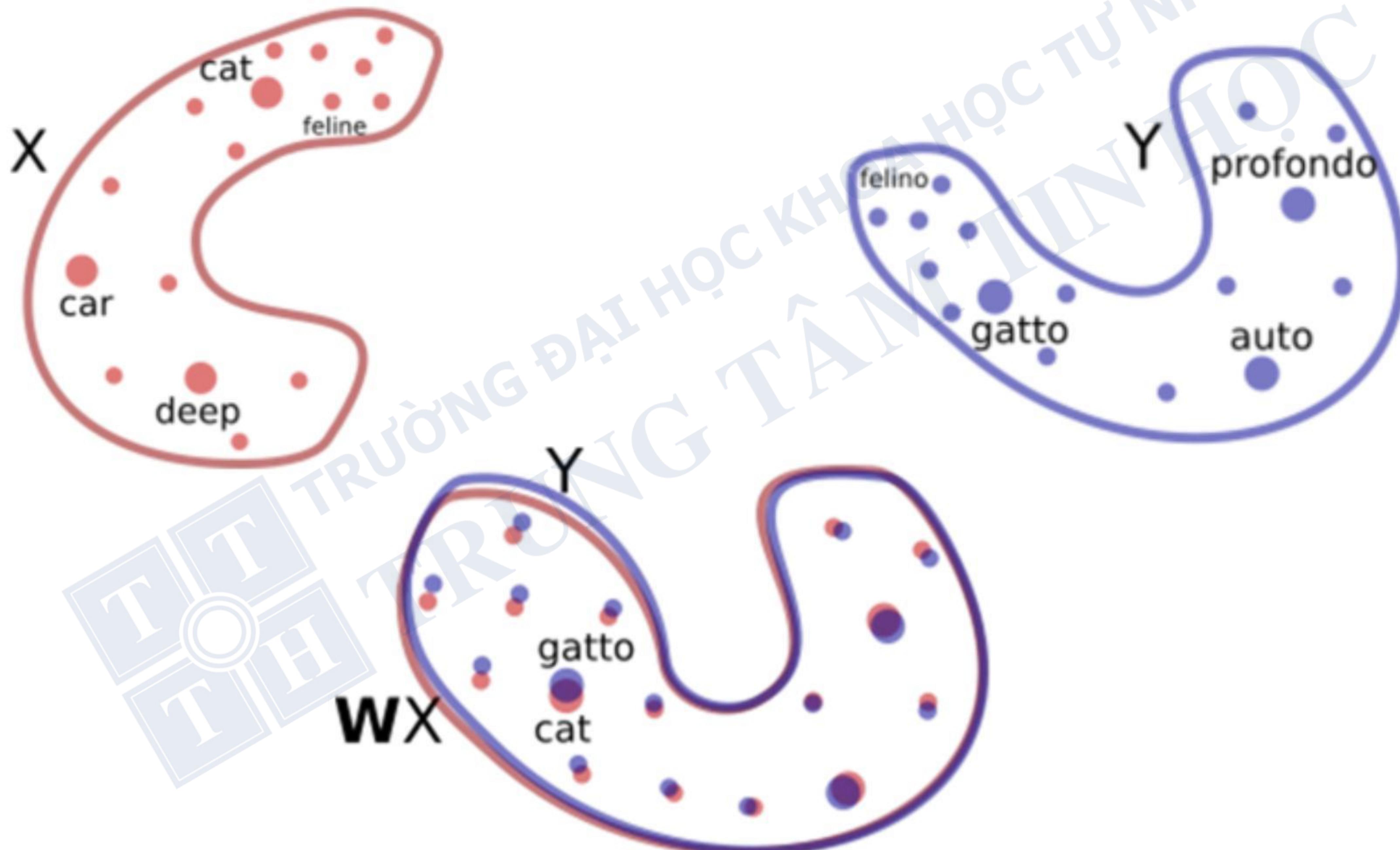
Tổng quan Machine Translation

K-nearest Neighbors



Tổng quan Machine Translation

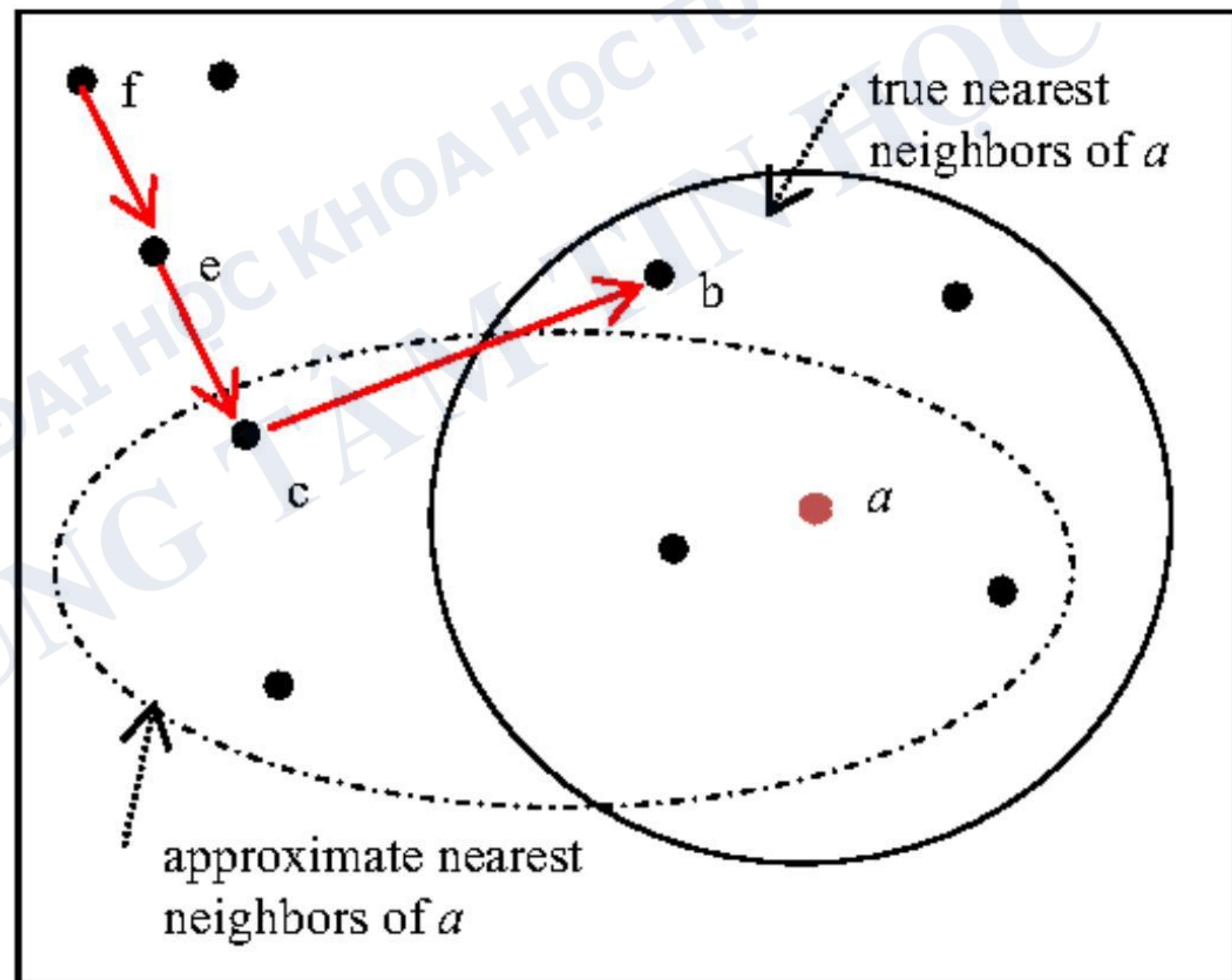
K-nearest Neighbors



Tổng quan Machine Translation

Approximate Nearest Neighbors

- Chỉ tính xấp xỉ.
- Đánh đổi độ chính xác để tăng hiệu suất.

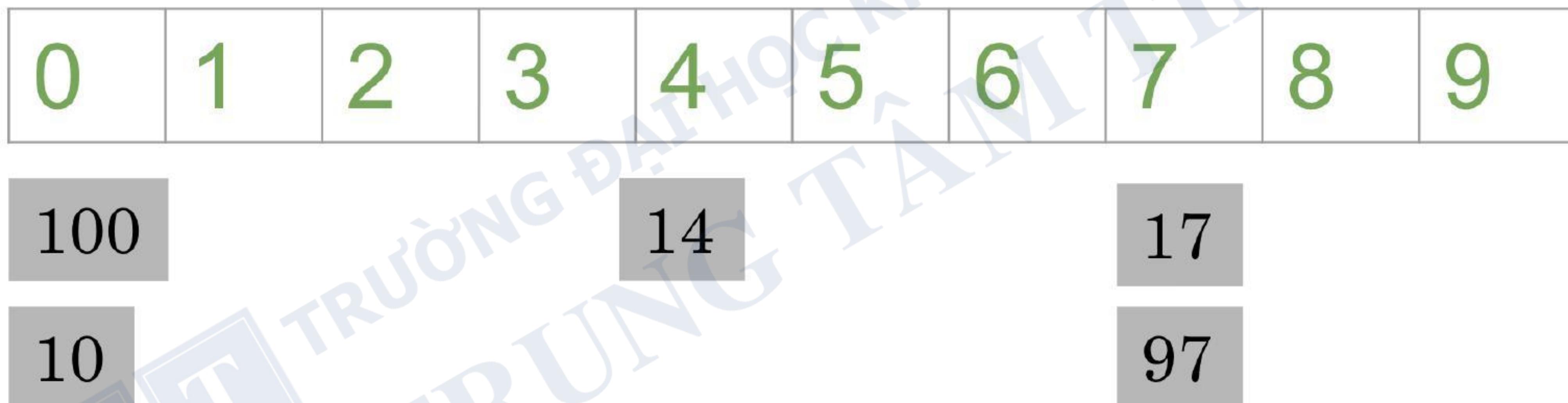




Tổng quan Machine Translation

Hash Functions

Là hàm lấy dữ liệu với kích thước tùy ý map với một không gian vector có kích thước cố định.



Hash function (vector) → Hash value

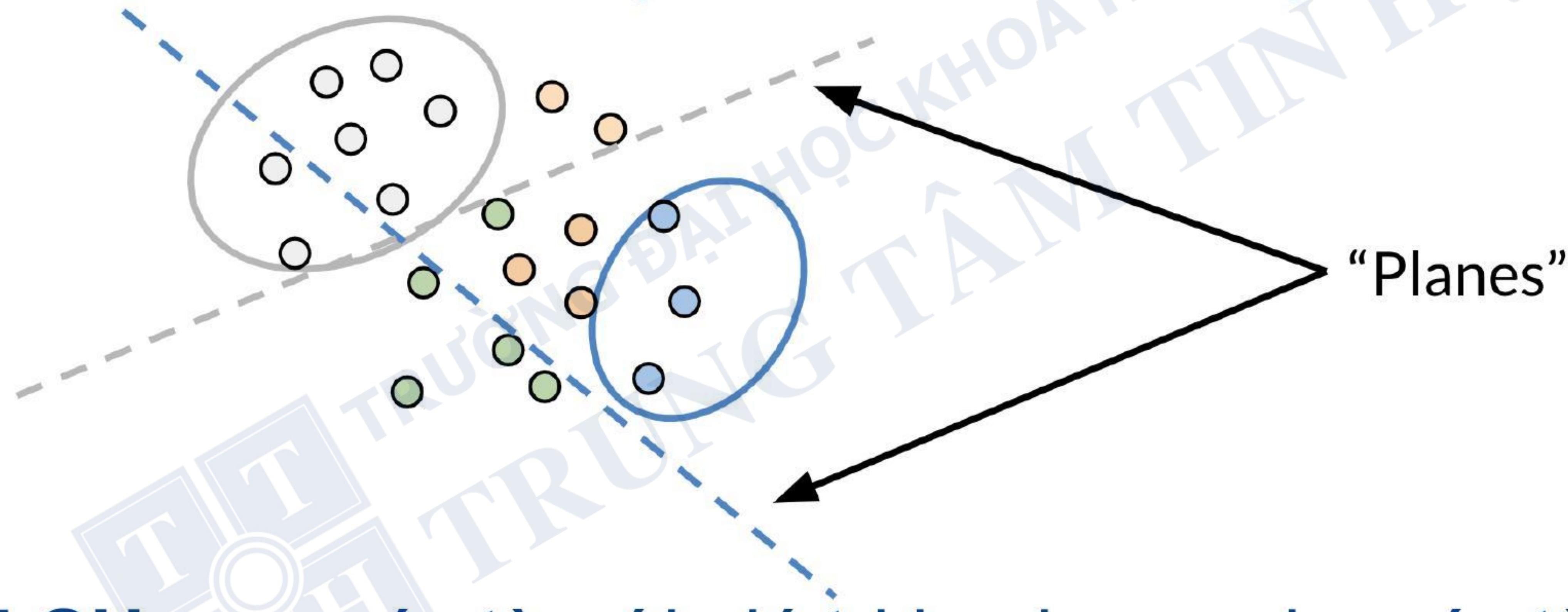
Hash value = vector % number of buckets



Tổng quan Machine Translation

Locality Sensitive Hashing - LSH

Là một phương pháp tìm kiếm và nhóm các từ dựa trên tính chất gần nhau của chúng.



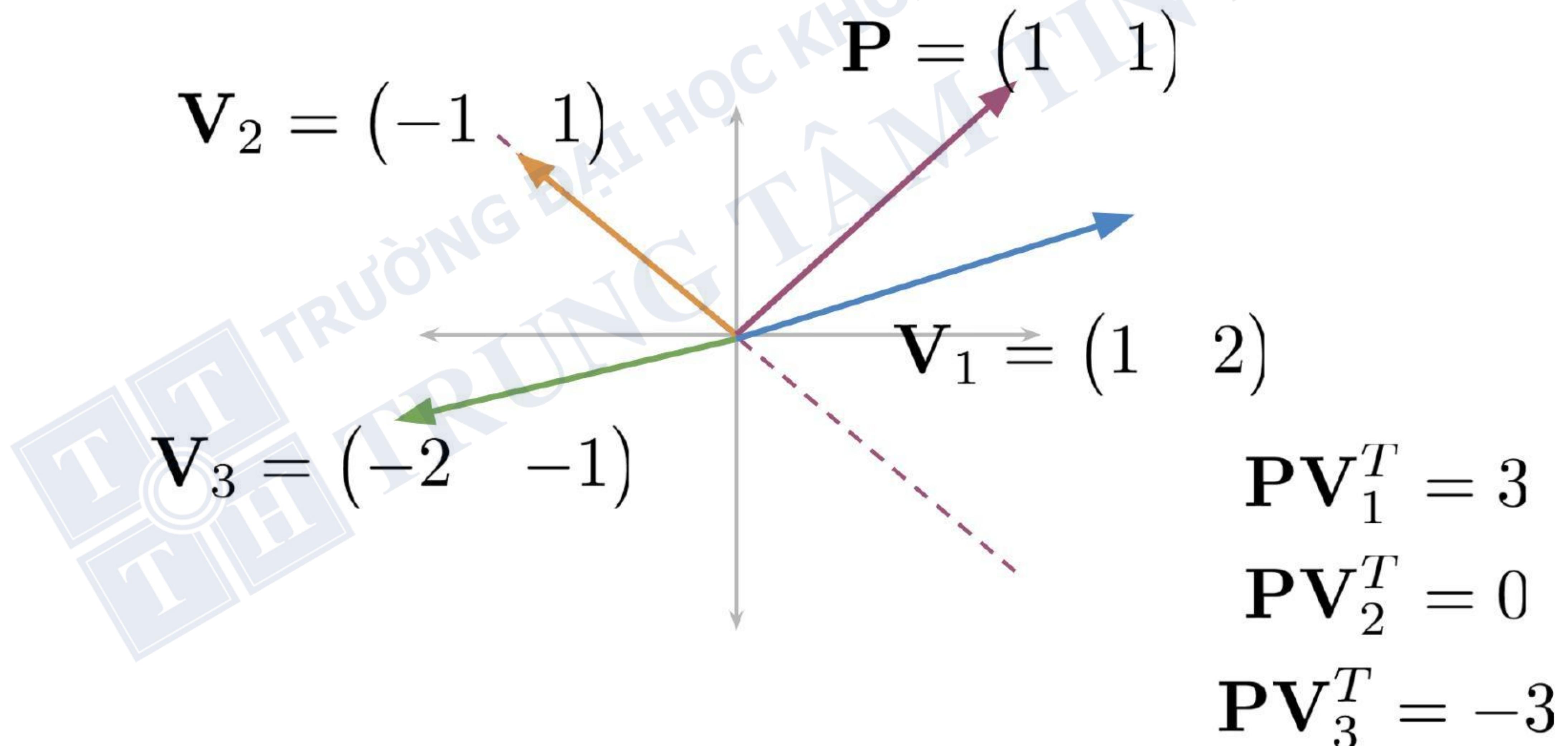
LSH map các từ với giá trị hash sao cho các từ có tính chất tương đồng có cùng giá trị hash hoặc có giá trị hash gần nhau.



Tổng quan Machine Translation

Locality Sensitive Hashing - LSH

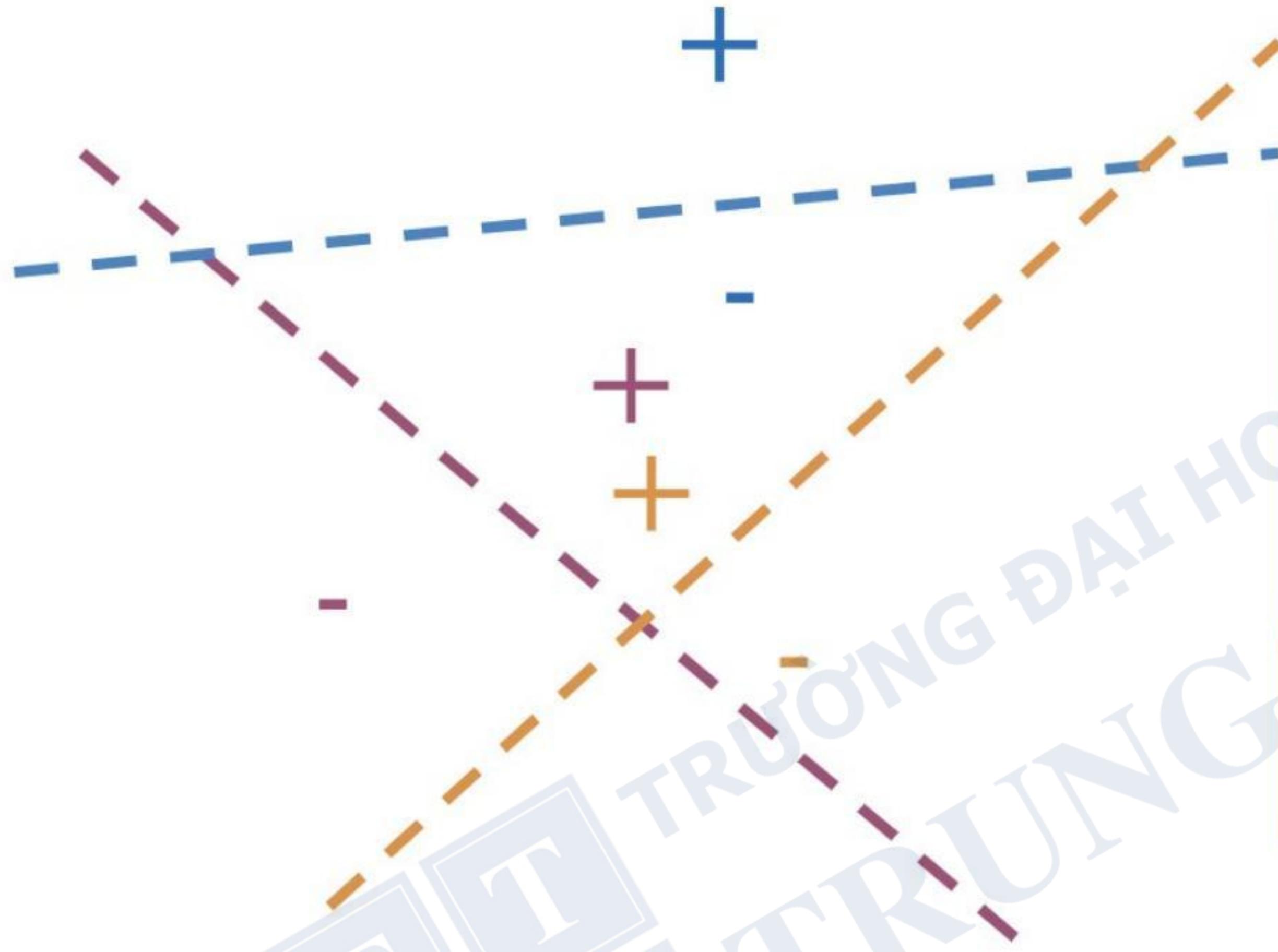
Là một phương pháp tìm kiếm và nhóm các từ dựa trên tính chất gần nhau của chúng.





Tổng quan Machine Translation

Locality Sensitive Hashing - LSH



$$\mathbf{P}_1 \mathbf{v}^T = 3, sign_1 = +1, h_1 = 1$$

$$\mathbf{P}_2 \mathbf{v}^T = 5, sign_2 = +1, h_2 = 1$$

$$\mathbf{P}_3 \mathbf{v}^T = -2, sign_3 = -1, h_3 = 0$$

Multiple Planes:
dùng nhiều plane để
tìm một giá trị hash.

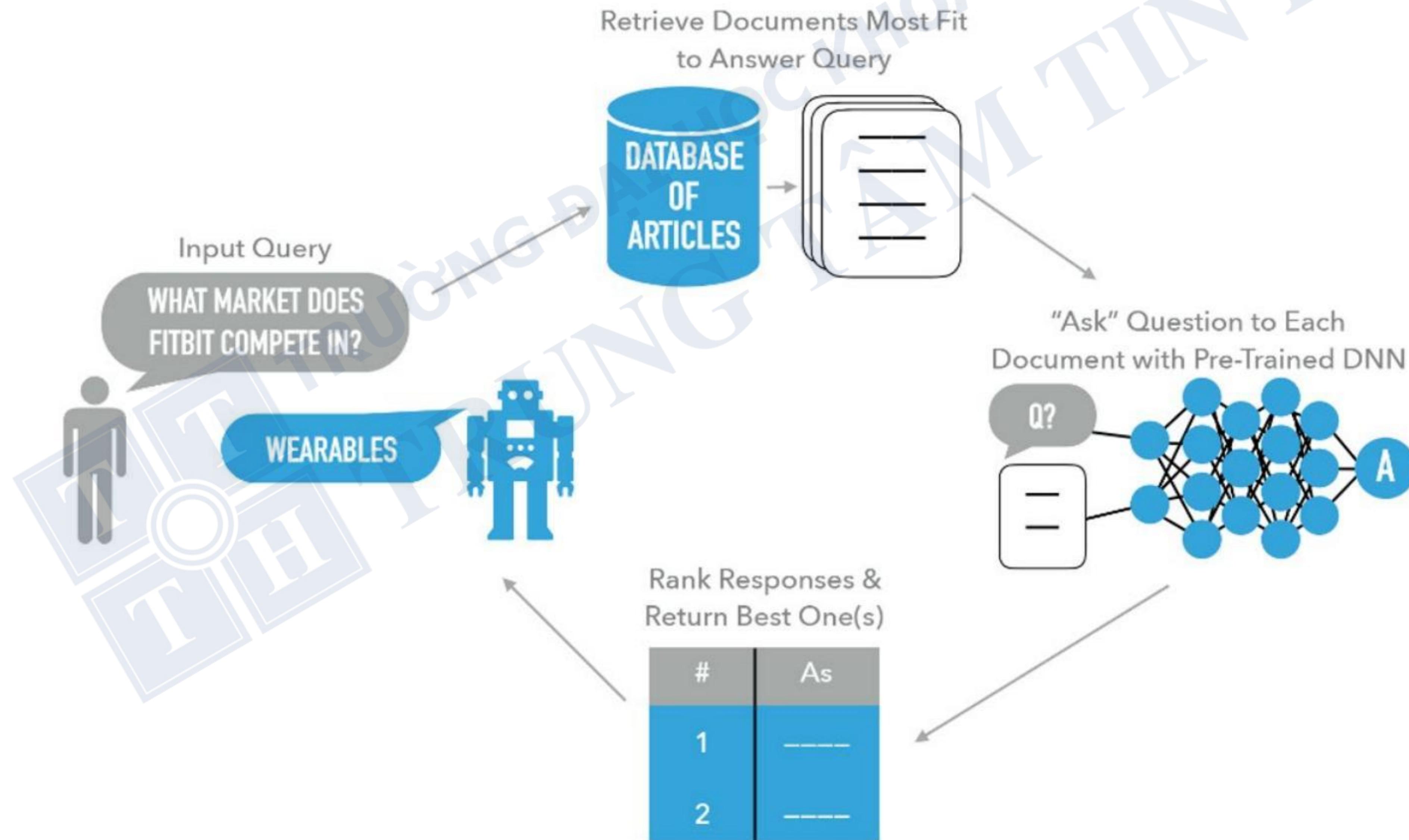
$$\begin{aligned}hash &= 2^0 \times h_1 + 2^1 \times h_2 + 2^2 \times h_3 \\&= 1 \times 1 + 2 \times 1 + 4 \times 0 \\&= 3\end{aligned}$$



Tổng quan Machine Translation

Document Search

Là quá trình tìm kiếm và truy xuất các văn bản dựa trên nội dung, từ khóa hay các tiêu chí khác.



MACHINE TRANSLATION



I. Tổng quan Machine Translation

II. Mô hình Seq2Seq

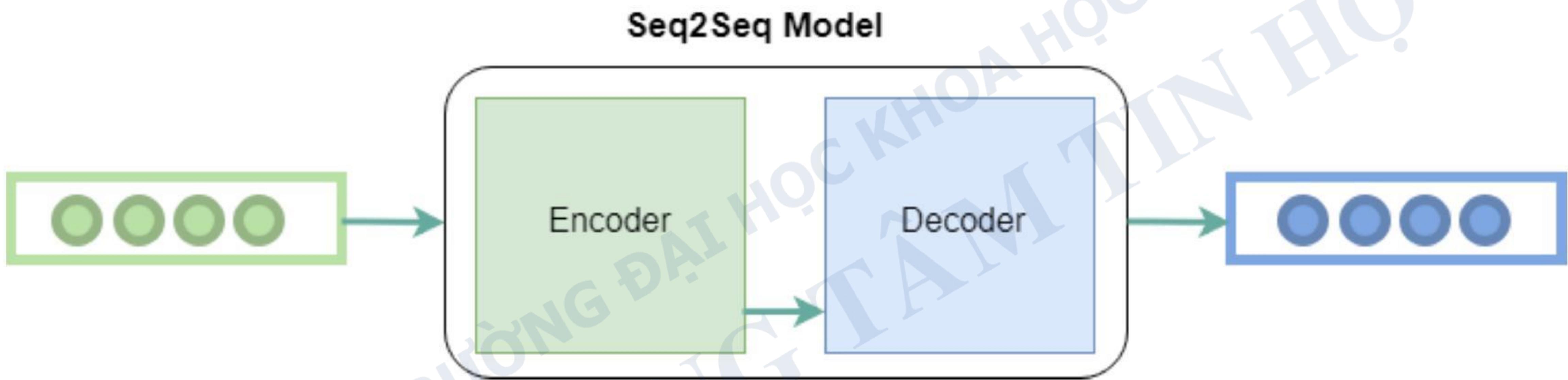
III. Mô hình Sep2seq + Attention

IV. Neural Machine Translation



Mô hình Seq2Seq

Sequence-to-Sequence sử dụng mạng neural để học và dự đoán từ dữ liệu chuỗi.

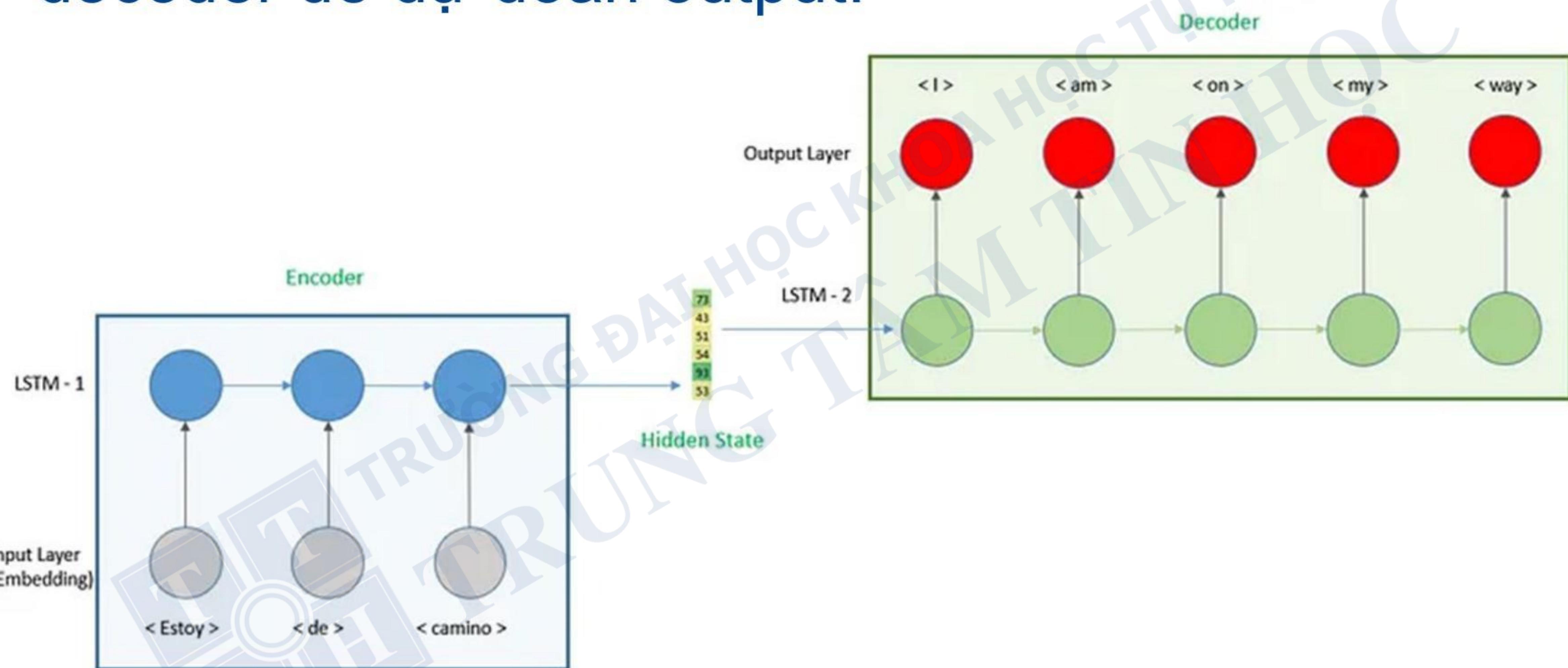


Encoder có input là chuỗi và map nó thành vector thể hiện ngữ cảnh.

Decoder nhận vector thể hiện ngữ cảnh từ encoder và dự đoán chuỗi output.

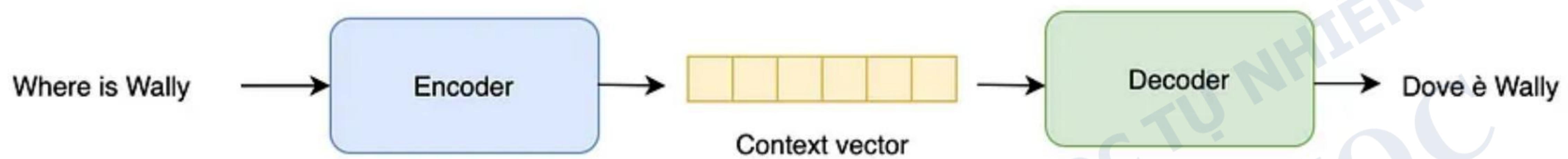
Mô hình Seq2Seq

Forward pass: input truyền qua encoder và decoder để dự đoán output.



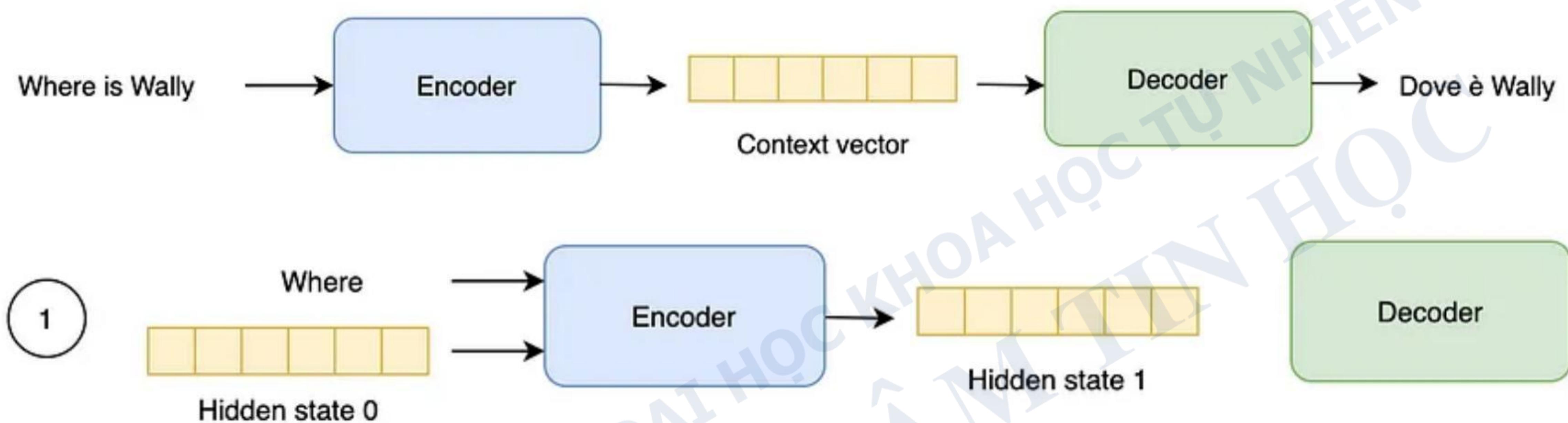
Backpropagation: cập nhật sai số giữa output dự đoán và thực tế và điều chỉnh mô hình.

Seq2Seq Encoder



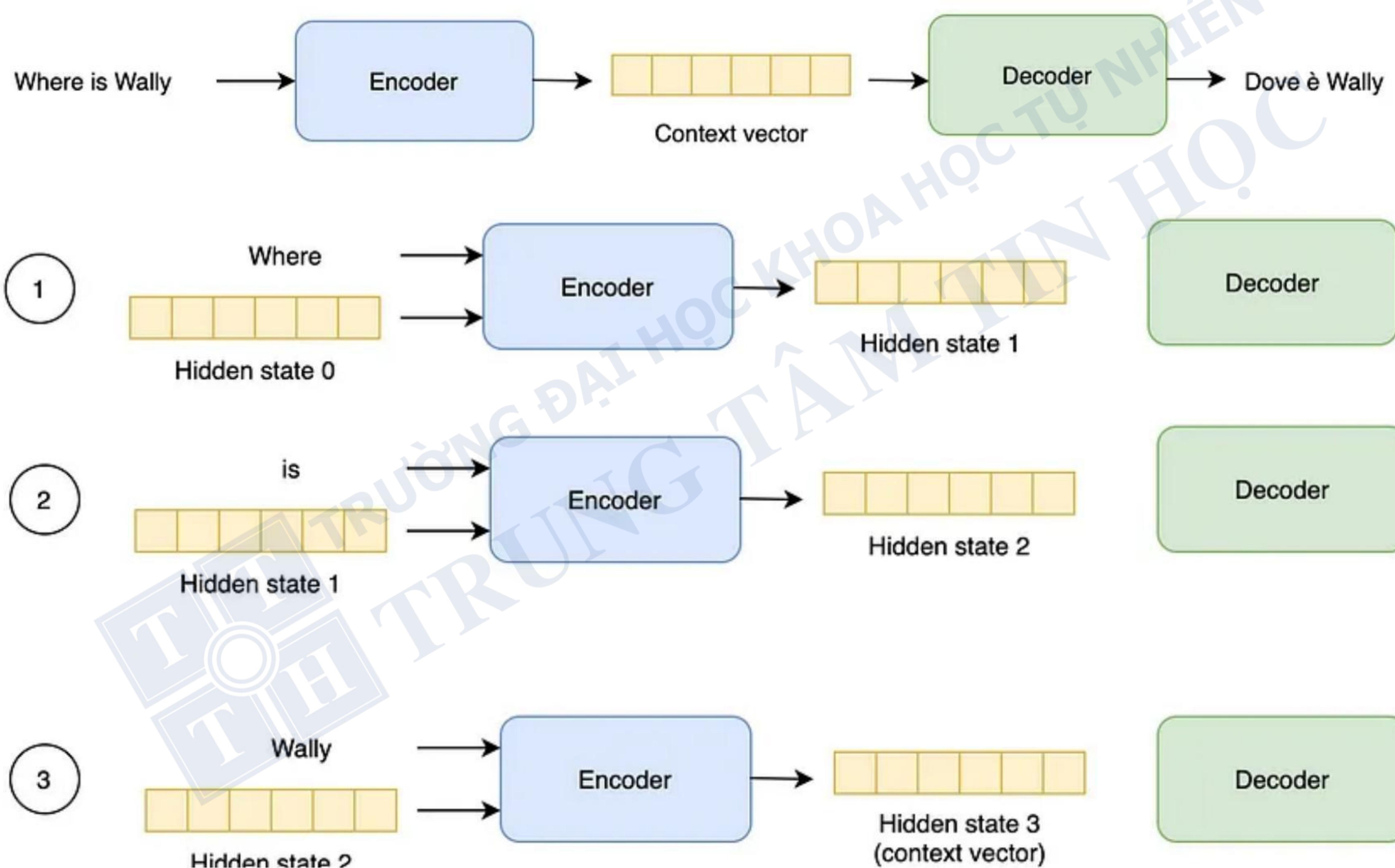


Seq2Seq Encoder





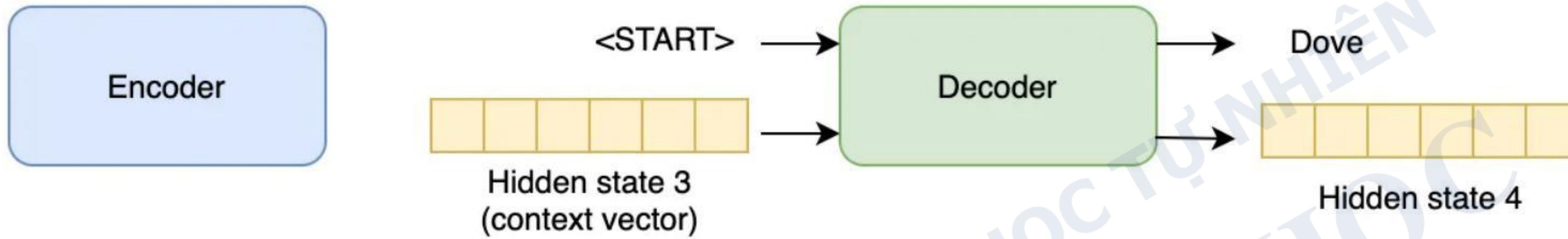
Seq2Seq Encoder

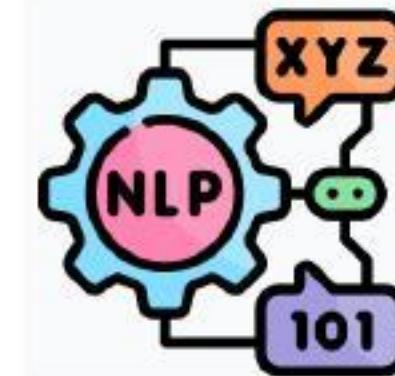




Seq2Seq Decoder

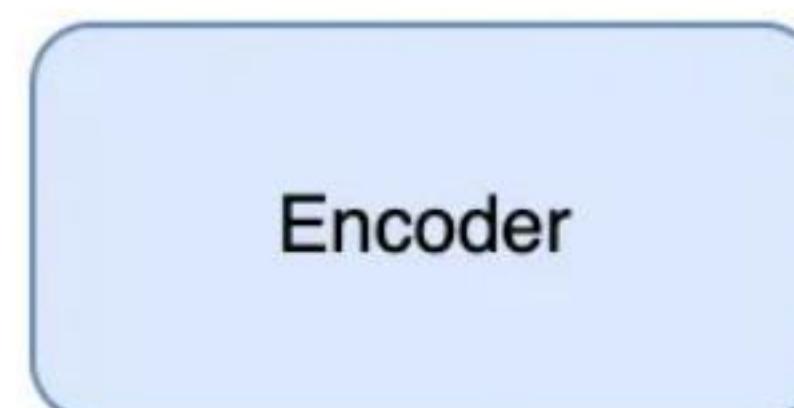
4



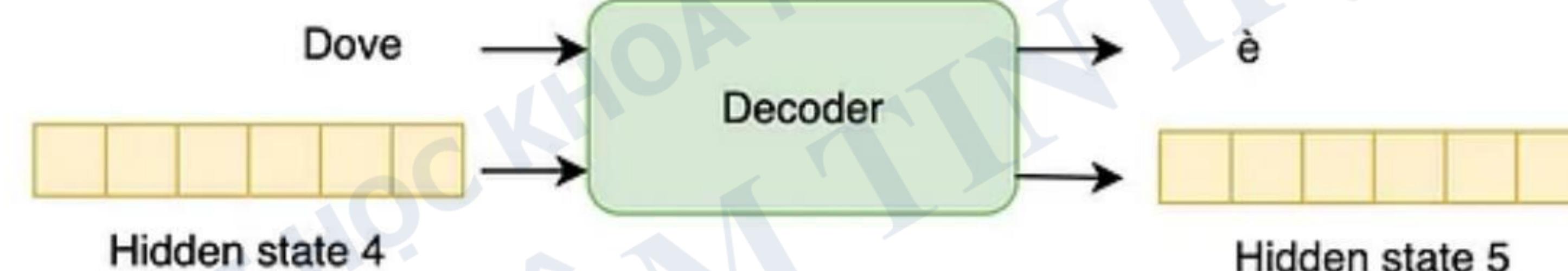
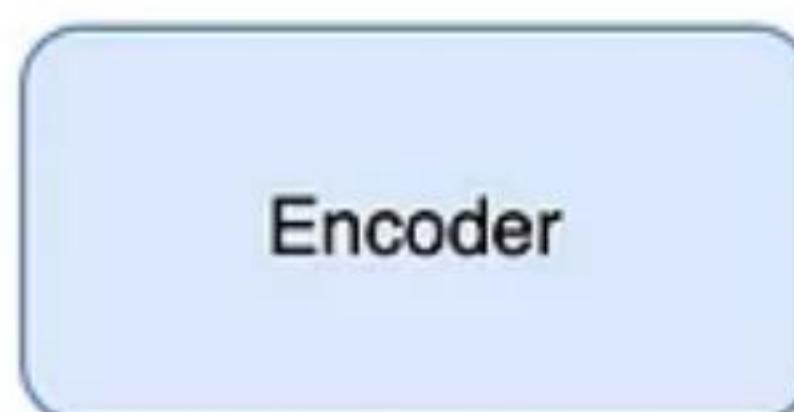


Seq2Seq Decoder

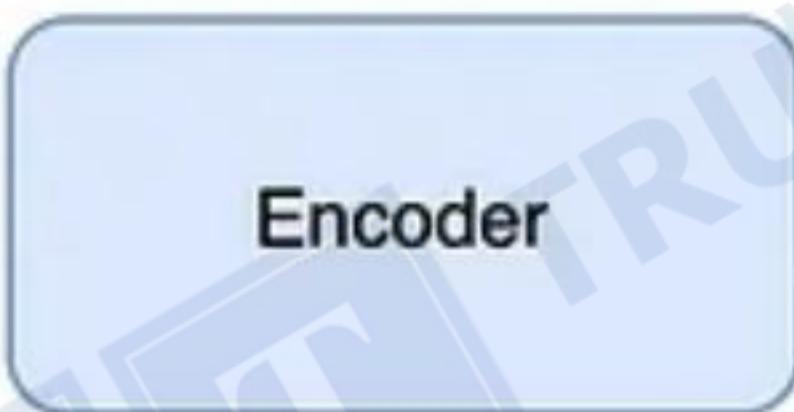
4



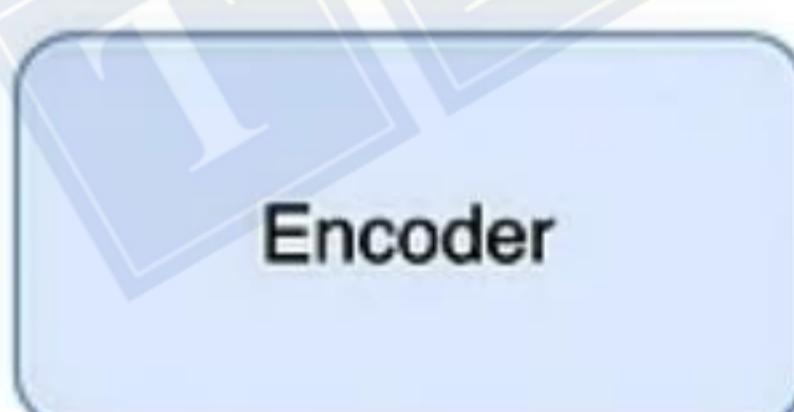
5



6

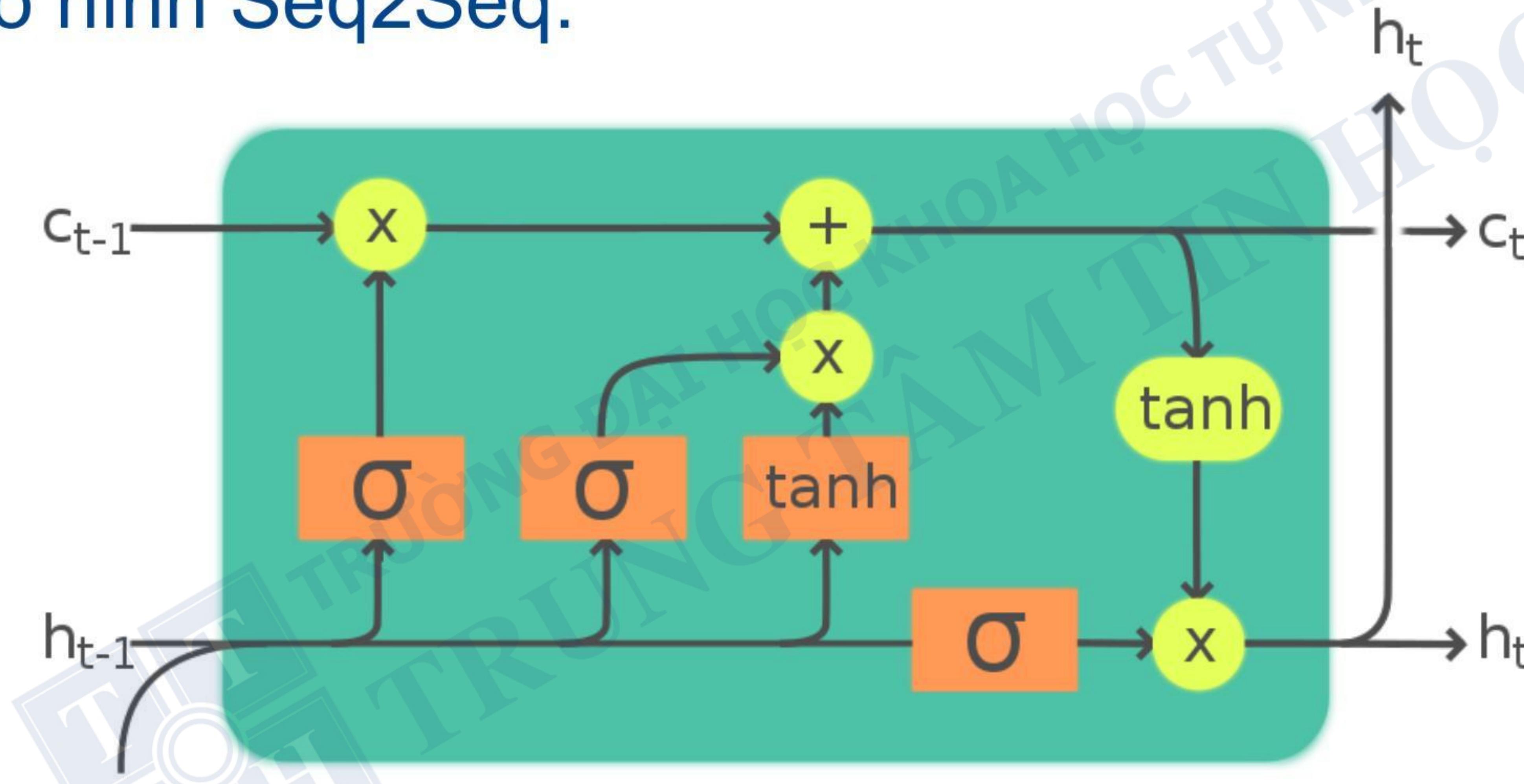


7



Mô hình Seq2Seq

Long Short-Term Memory (LSTM) là một dạng mô hình Seq2Seq.

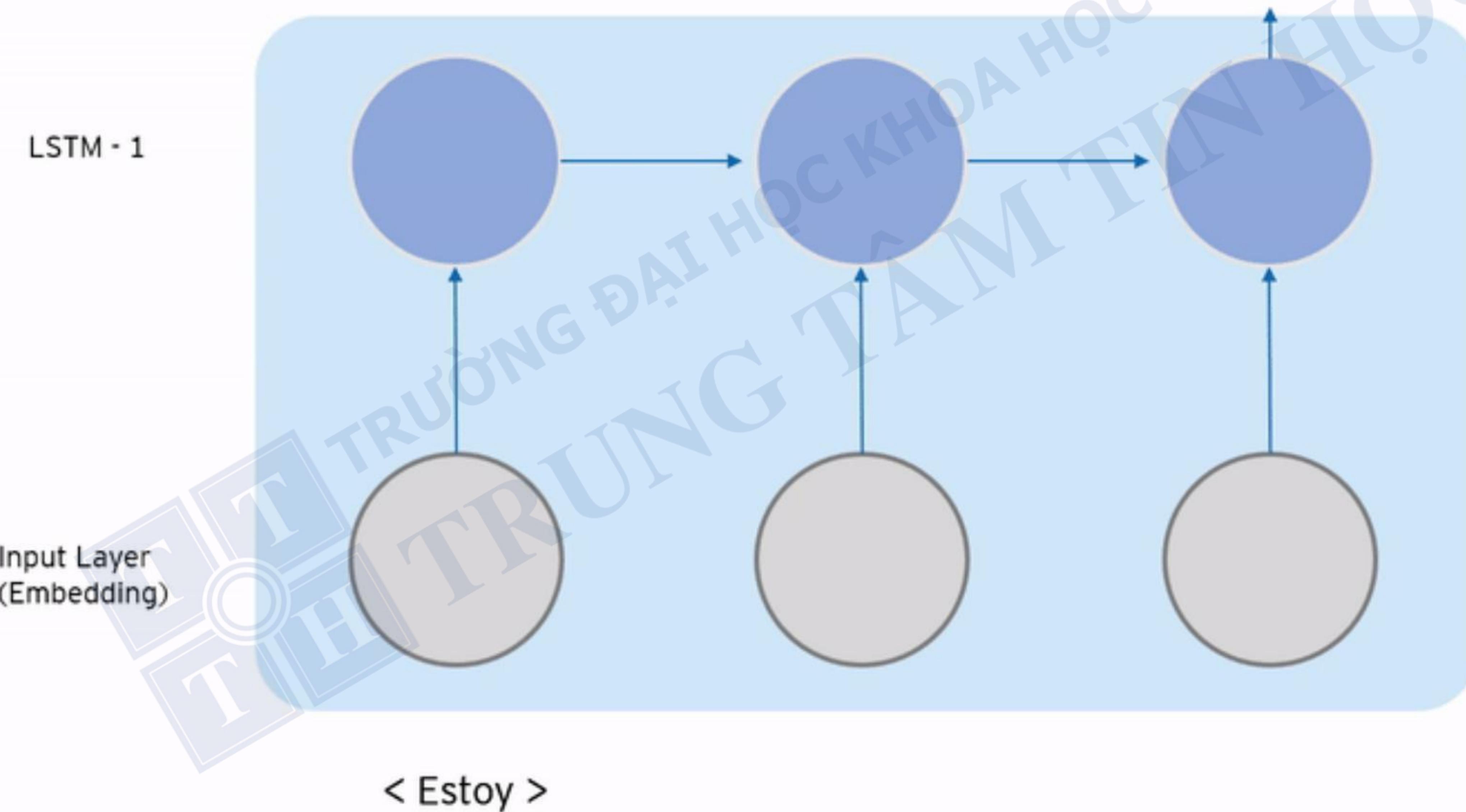


Legend:

Layer	ComponentwiseCopy	Concatenate

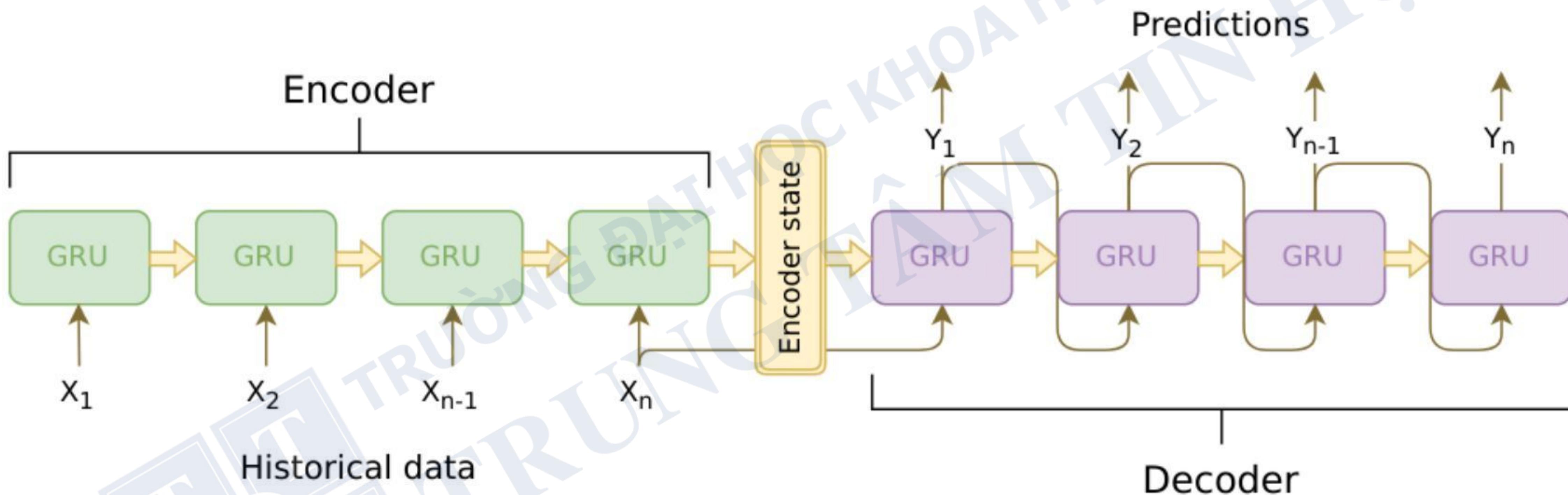


Mô hình Seq2Seq



Ưu điểm mô hình Seq2Seq

- Xử lý các tác vụ liên quan đến chuỗi input và chuỗi output.
- Học và hiểu các mối quan hệ phức tạp giữa các từ và câu.
- Tự động học từ dữ liệu và cải thiện hiệu suất theo thời gian.



Seq2Seq sử dụng để xử lý các tác vụ liên quan đến chuỗi trong NLP: machine translation, chatbot, sentiment analysis, trích xuất thông tin.

MACHINE TRANSLATION

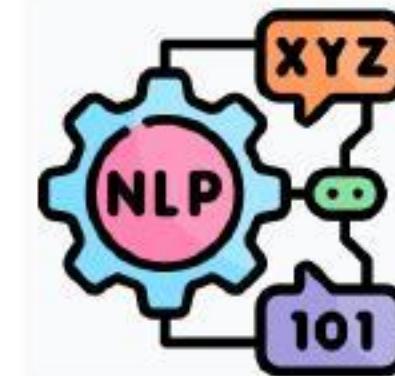


I. Tổng quan Machine Translation

II. Mô hình Seq2Seq

III. Mô hình Seq2seq + Attention

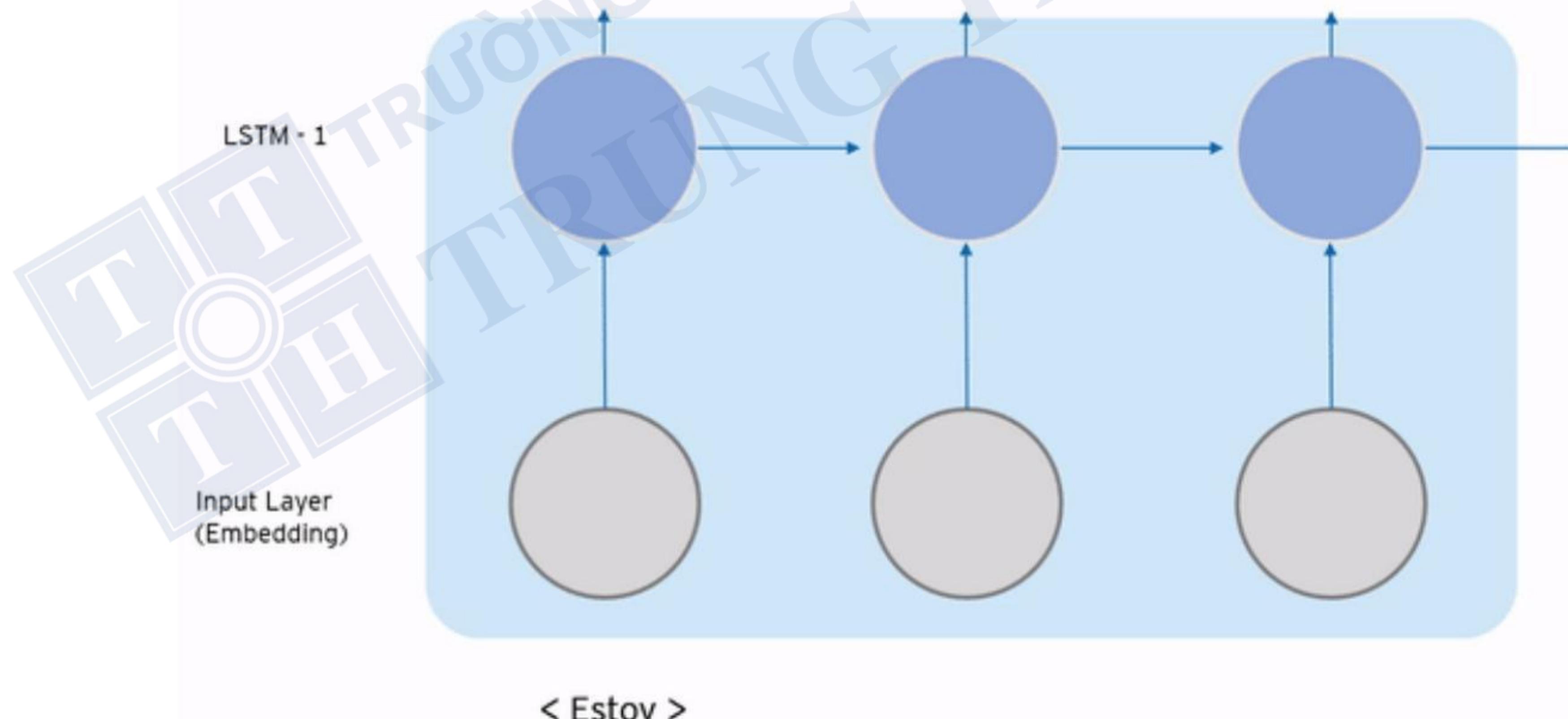
IV. Neural Machine Translation



Mô hình Seq2Seq + Attention

Seq2Seq+Attention là một biến thể của mô hình Seq2Seq.

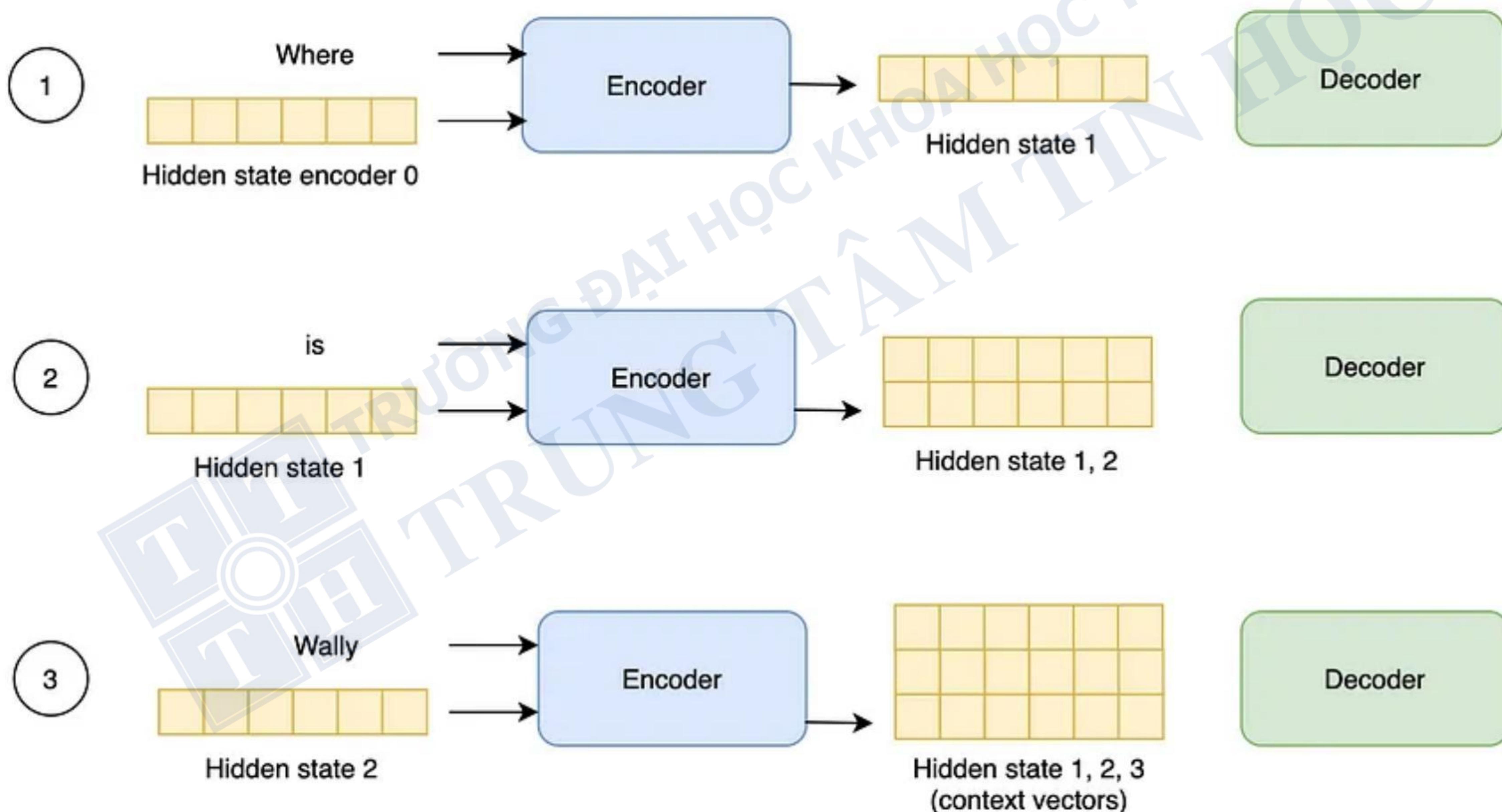
→ Giải quyết vấn đề của mô hình Seq2Seq khi xử lý các chuỗi dài hay có sự phụ thuộc xa giữa các từ (**bottleneck problem**).





Mô hình Seq2Seq + Attention

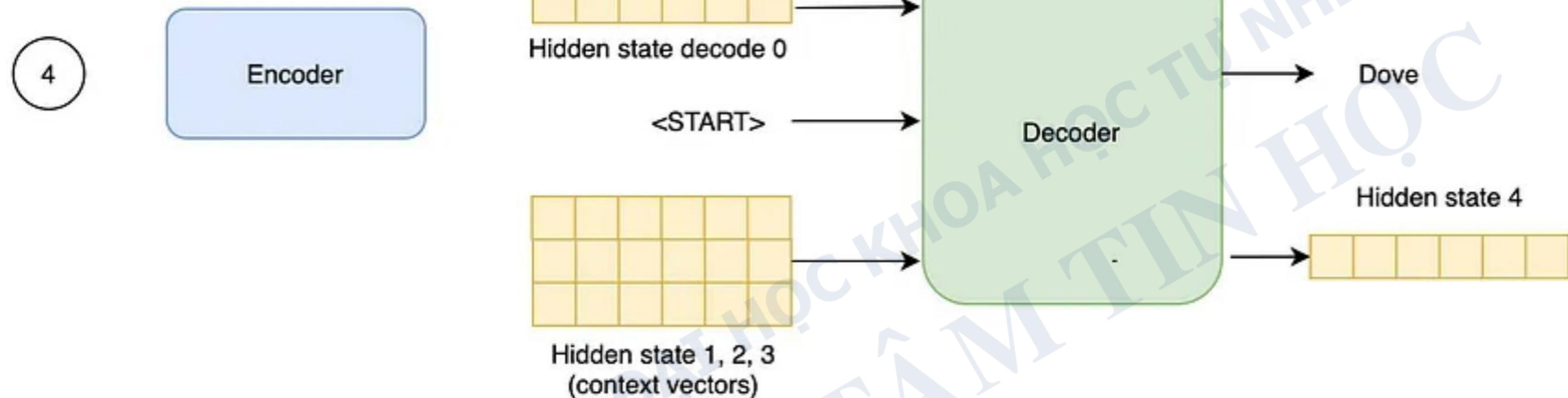
Encoder tạo **chuỗi các vector** biểu diễn ngũ cảnh, mỗi vector tương ứng với một từ trong câu đầu vào.



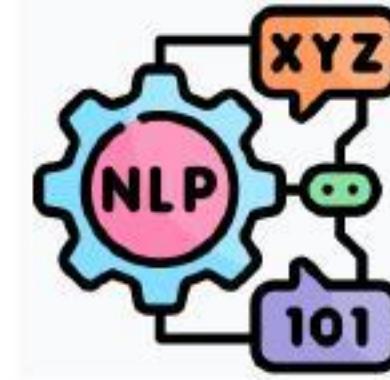
Mô hình Seq2Seq + Attention



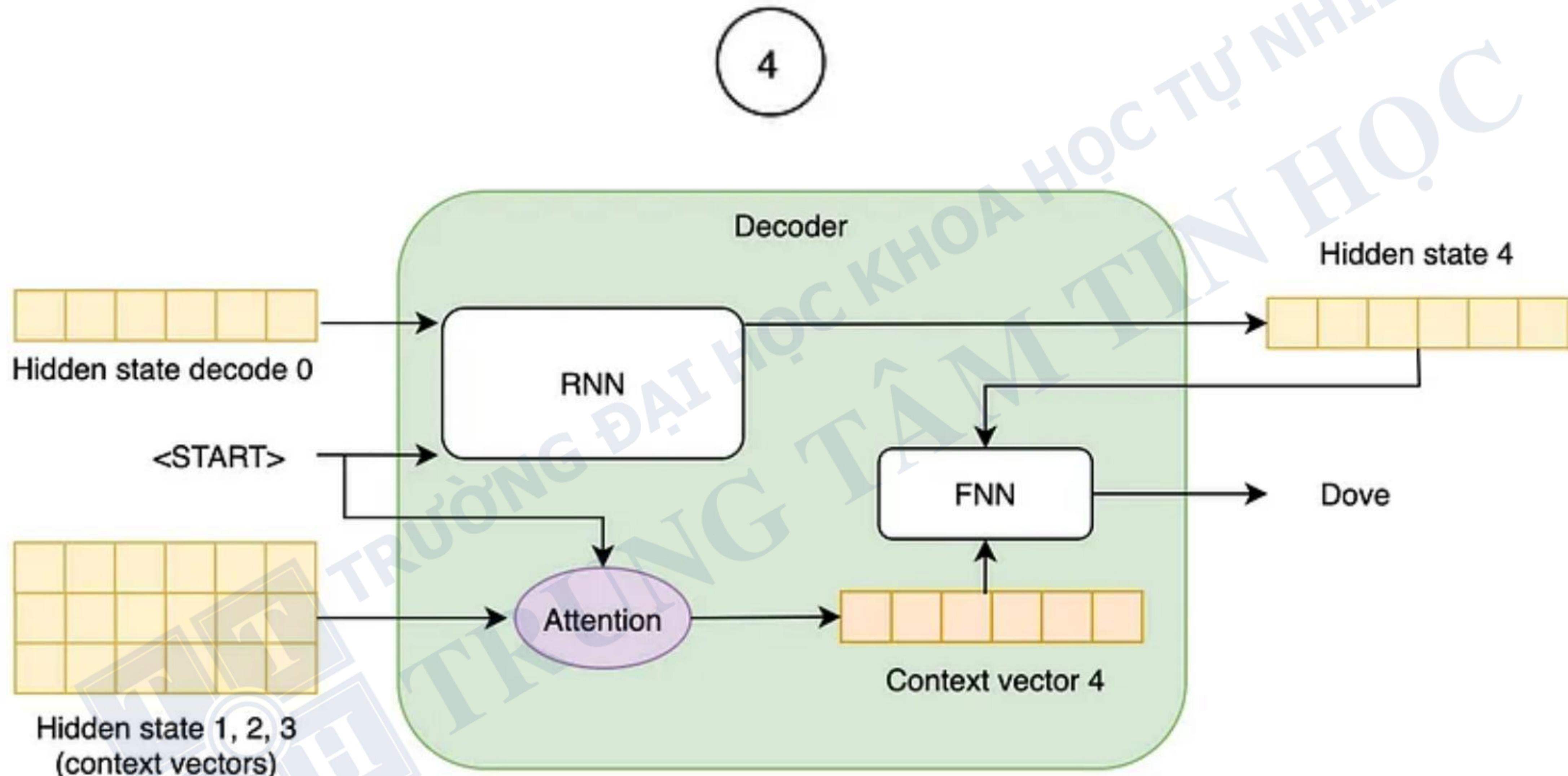
Decoder:



Mô hình Seq2Seq + Attention



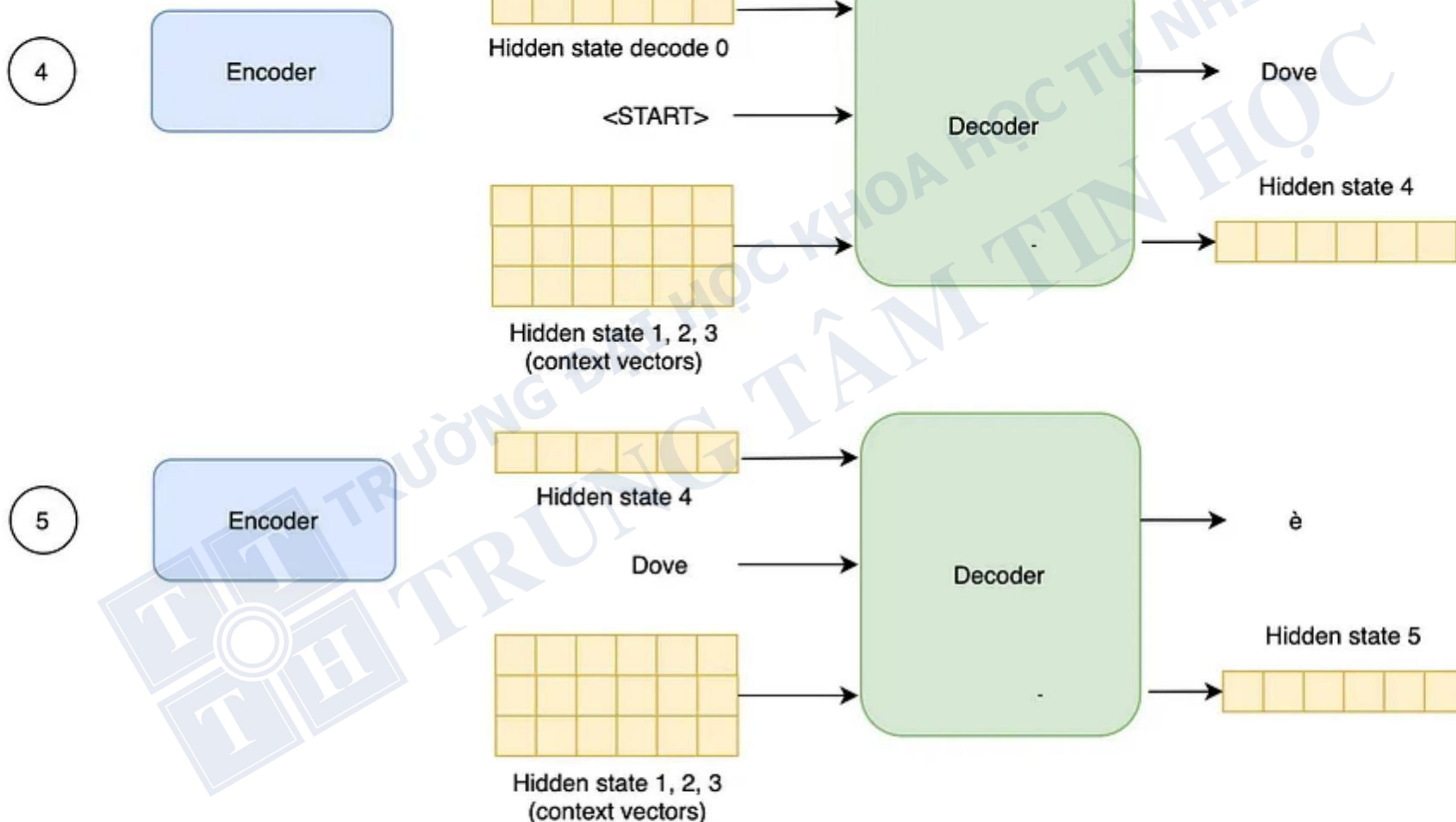
Decoder:





Mô hình Seq2Seq + Attention

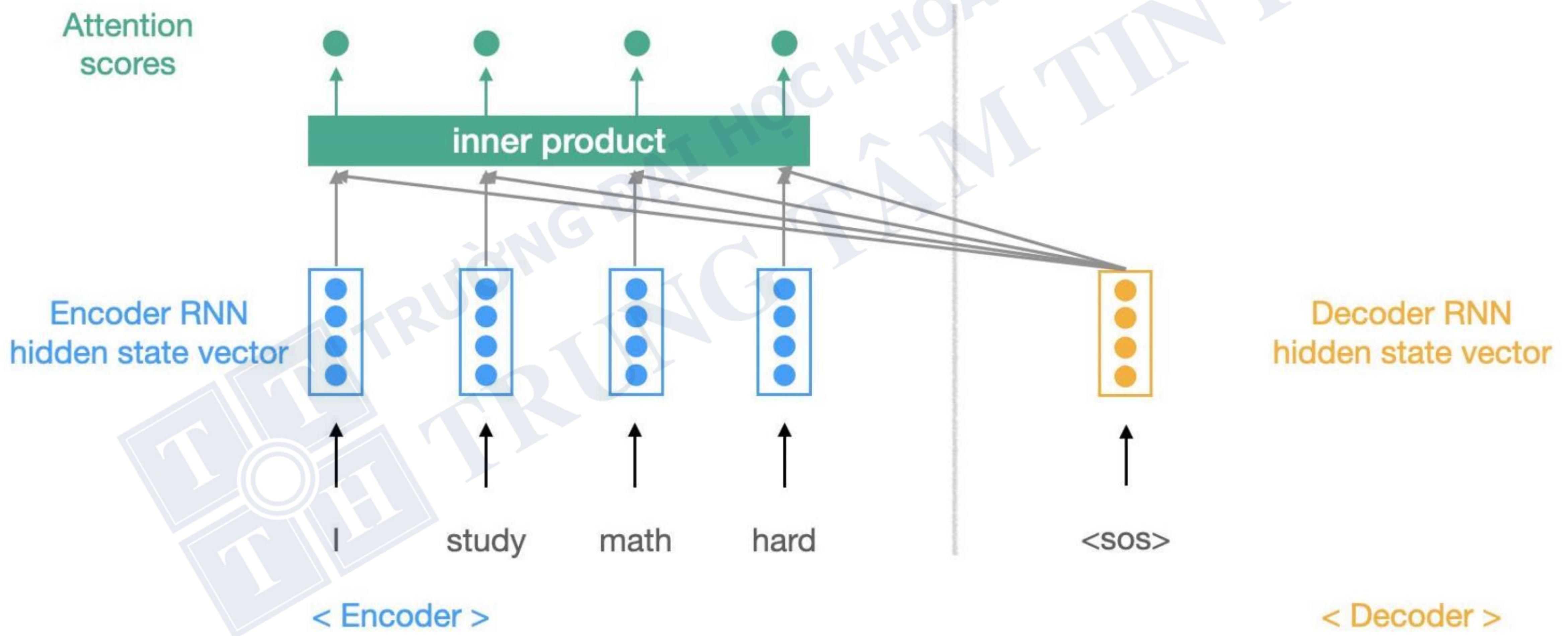
Decoder:



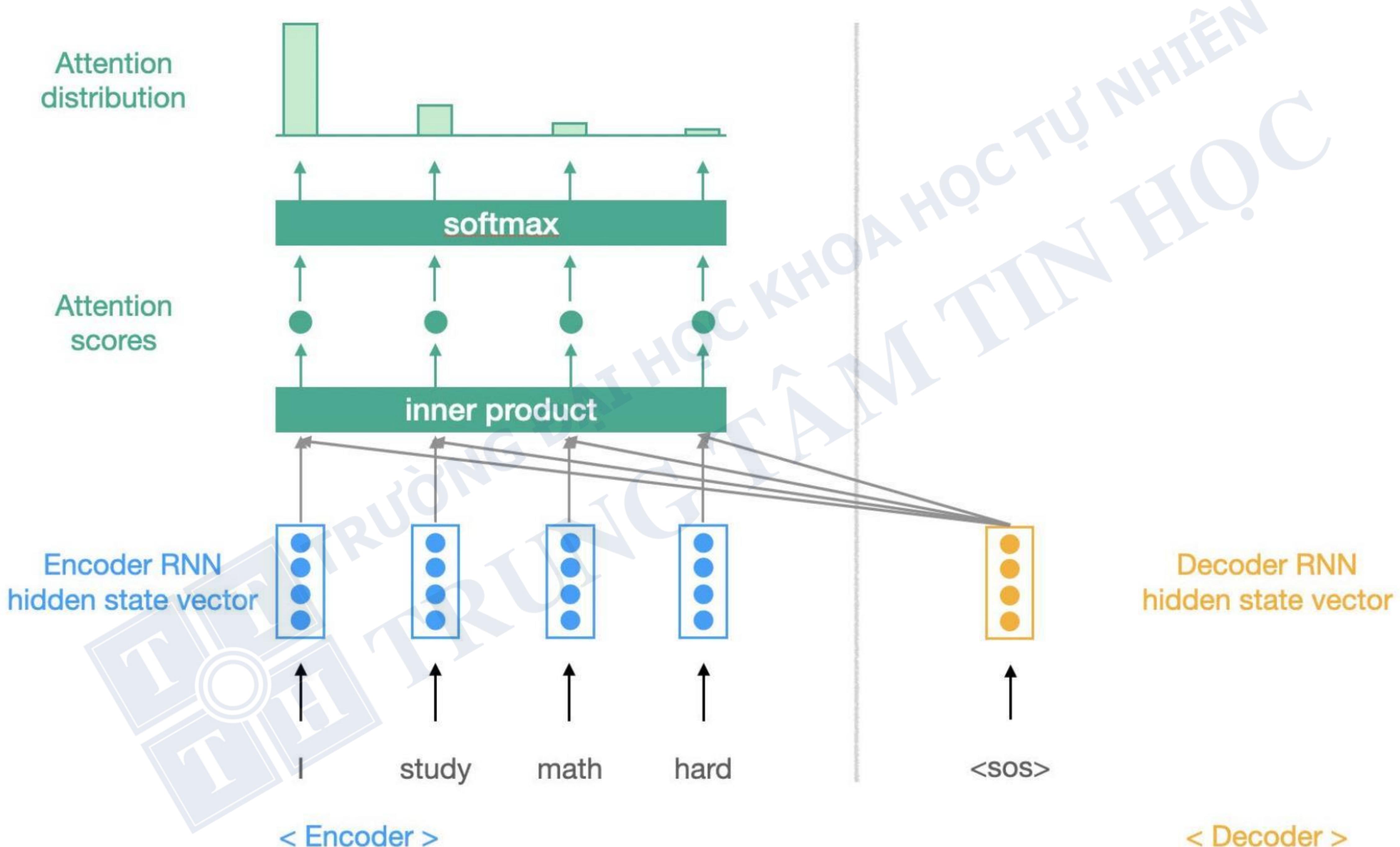


Mô hình Seq2Seq + Attention

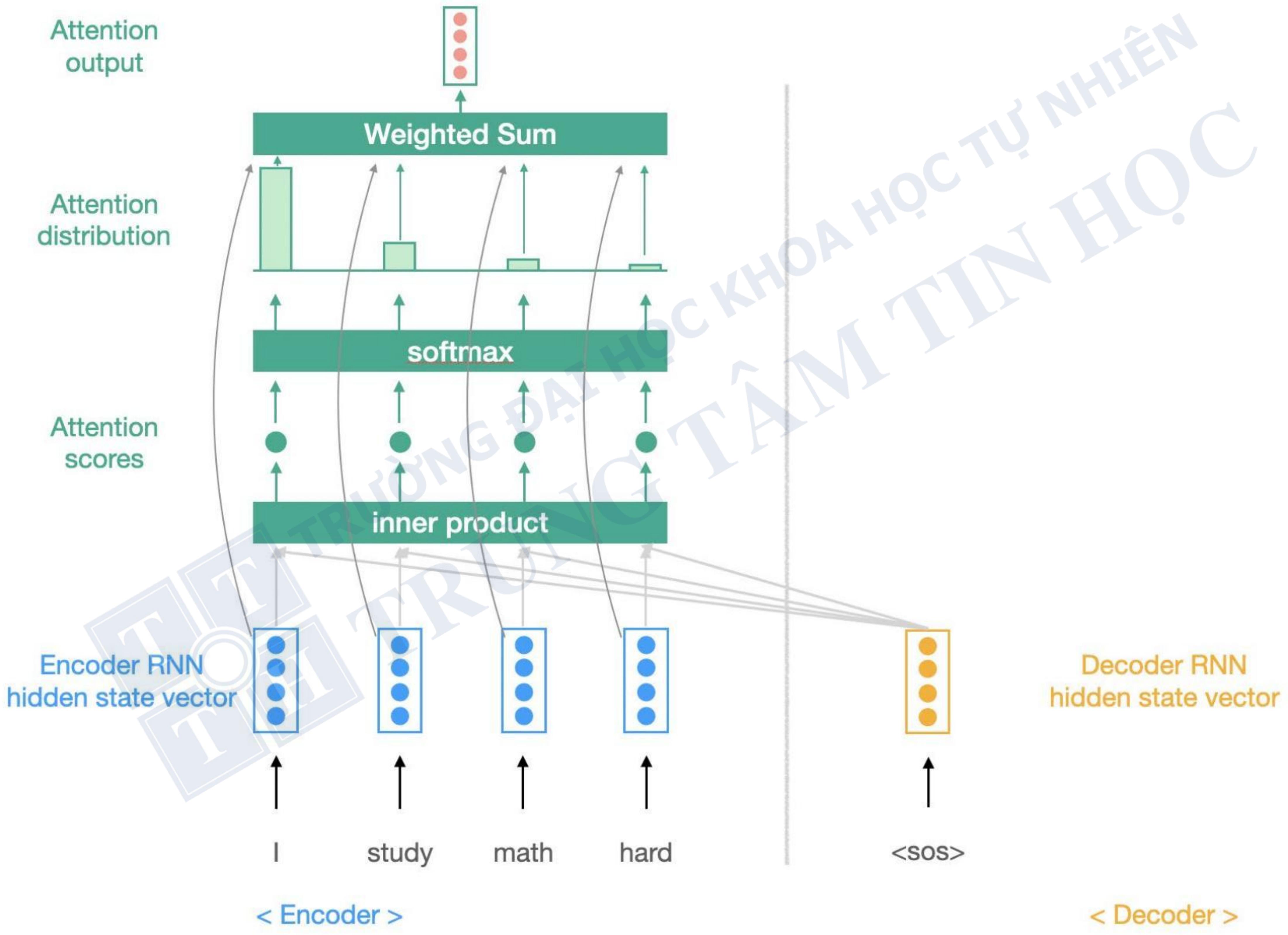
Seq2Seq+Attention là một biến thể của mô hình Seq2Seq.



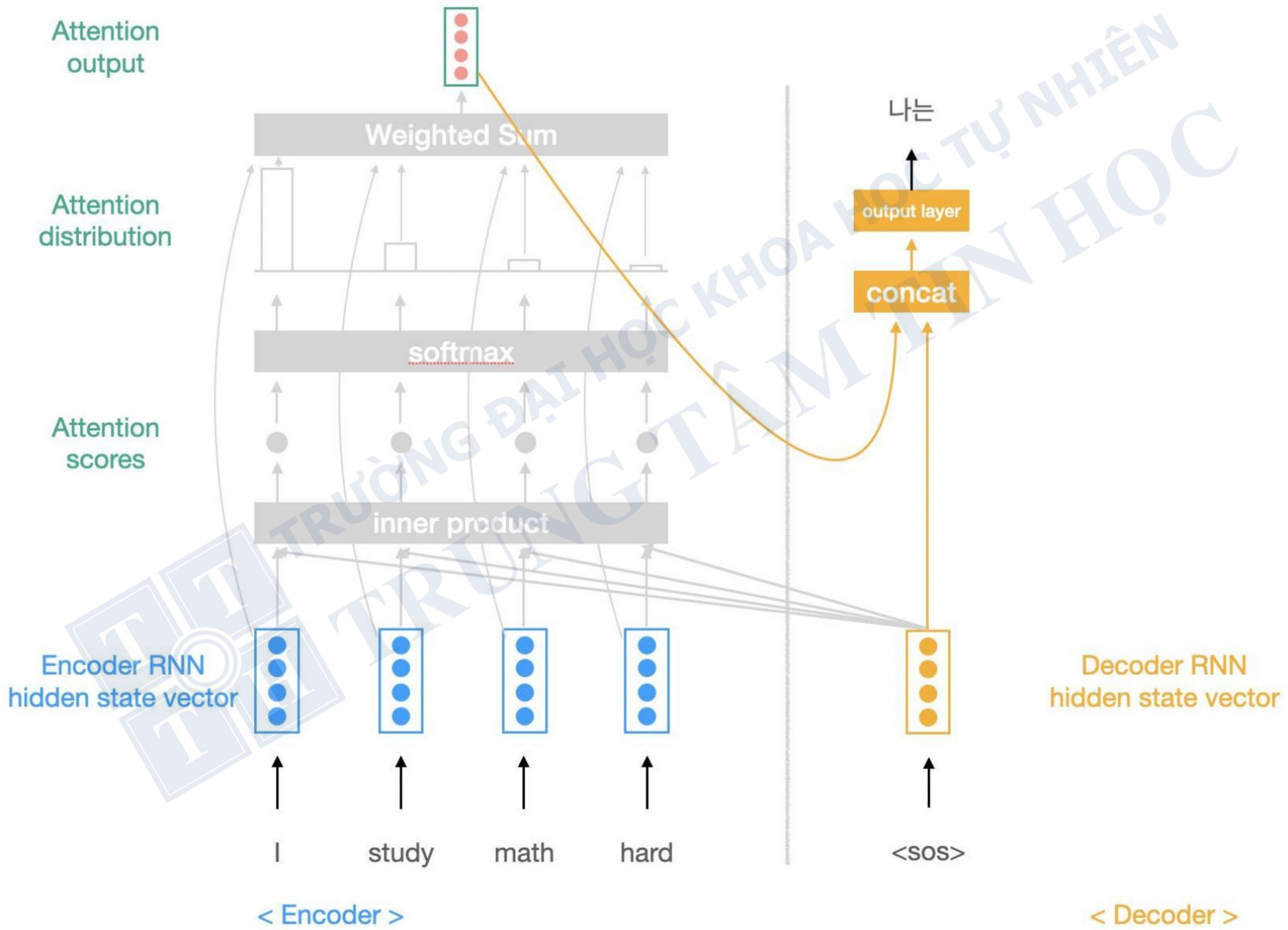
Mô hình Seq2Seq + Attention



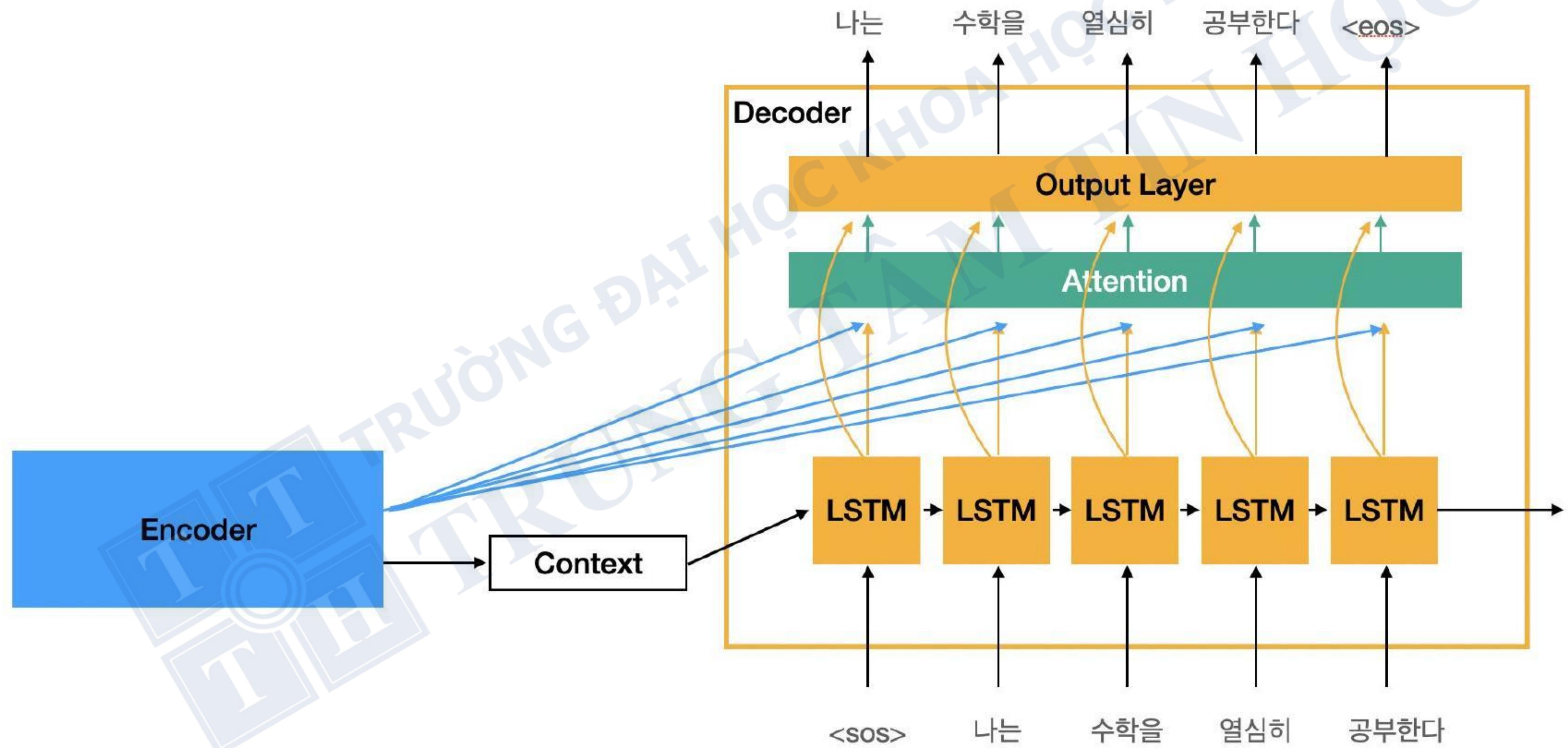
Mô hình Seq2Seq + Attention



Mô hình Seq2Seq + Attention



Mô hình Seq2Seq + Attention





Mô hình Seq2Seq + Attention

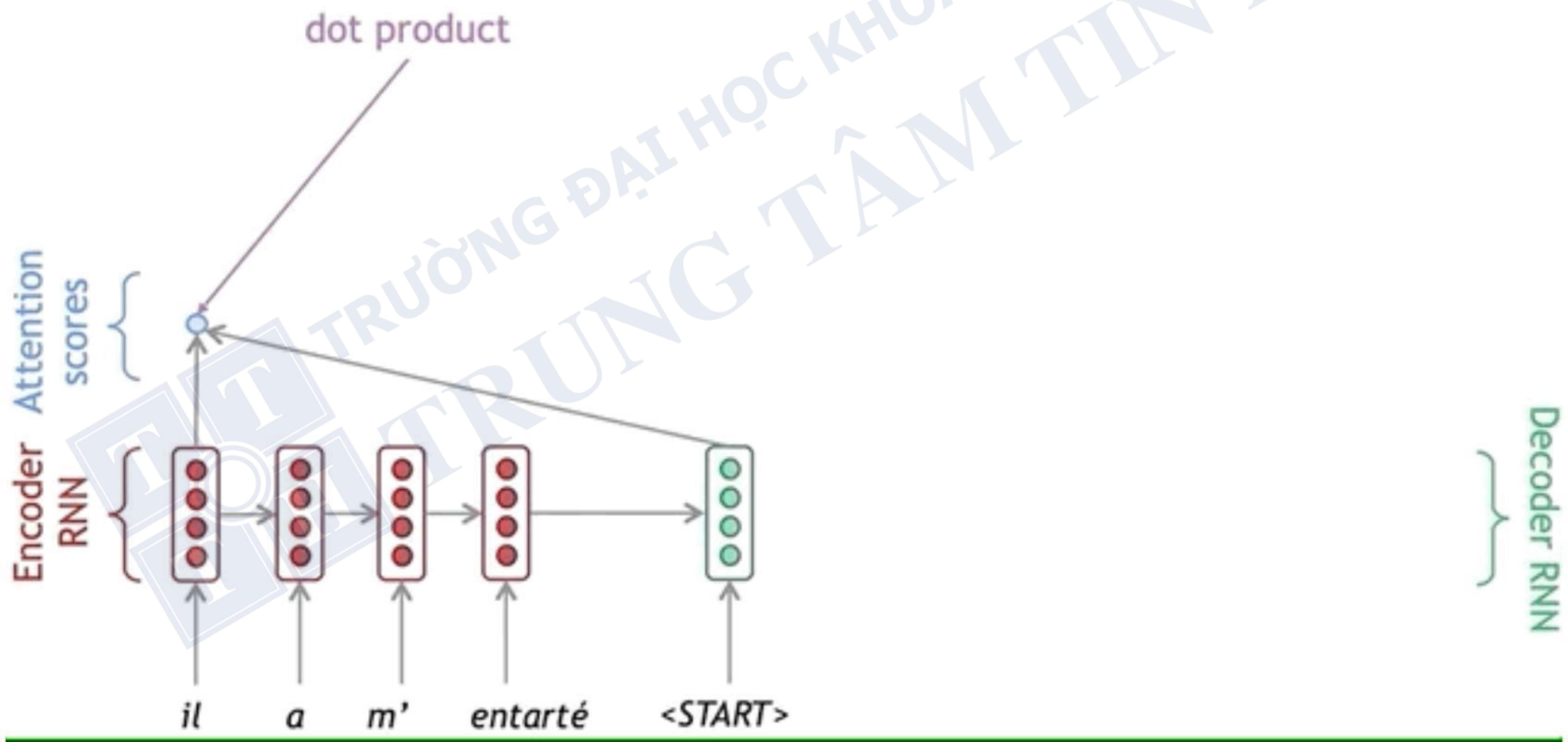
. la maison de Léa <end>





Mô hình Seq2Seq + Attention

Xử lý các tác vụ liên quan đến chuỗi input và chuỗi output.
Học và hiểu các mối quan hệ phức tạp giữa các thời gian.



MACHINE TRANSLATION



I. Tổng quan Machine Translation

II. Mô hình Seq2Seq

III. Mô hình Seq2seq + Attention

IV. Neural Machine Translation

Neural Machine Translation



Là mô hình sử dụng **neural network** để học và dự đoán chuỗi để dịch ngôn ngữ trong NLP.

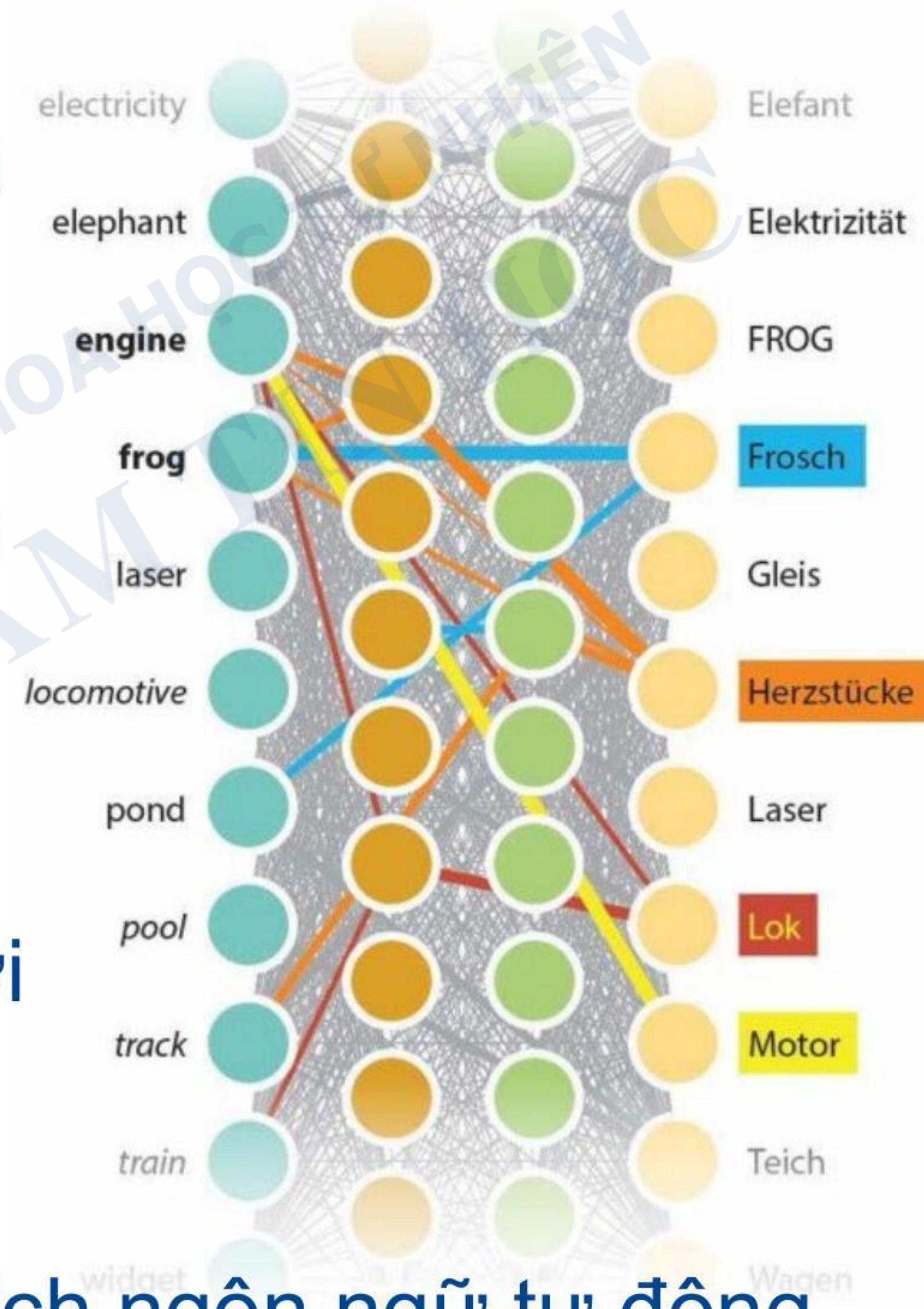
Neural machine translation sử dụng bộ encoder-decoder kết hợp **attention**.



Neural Machine Translation

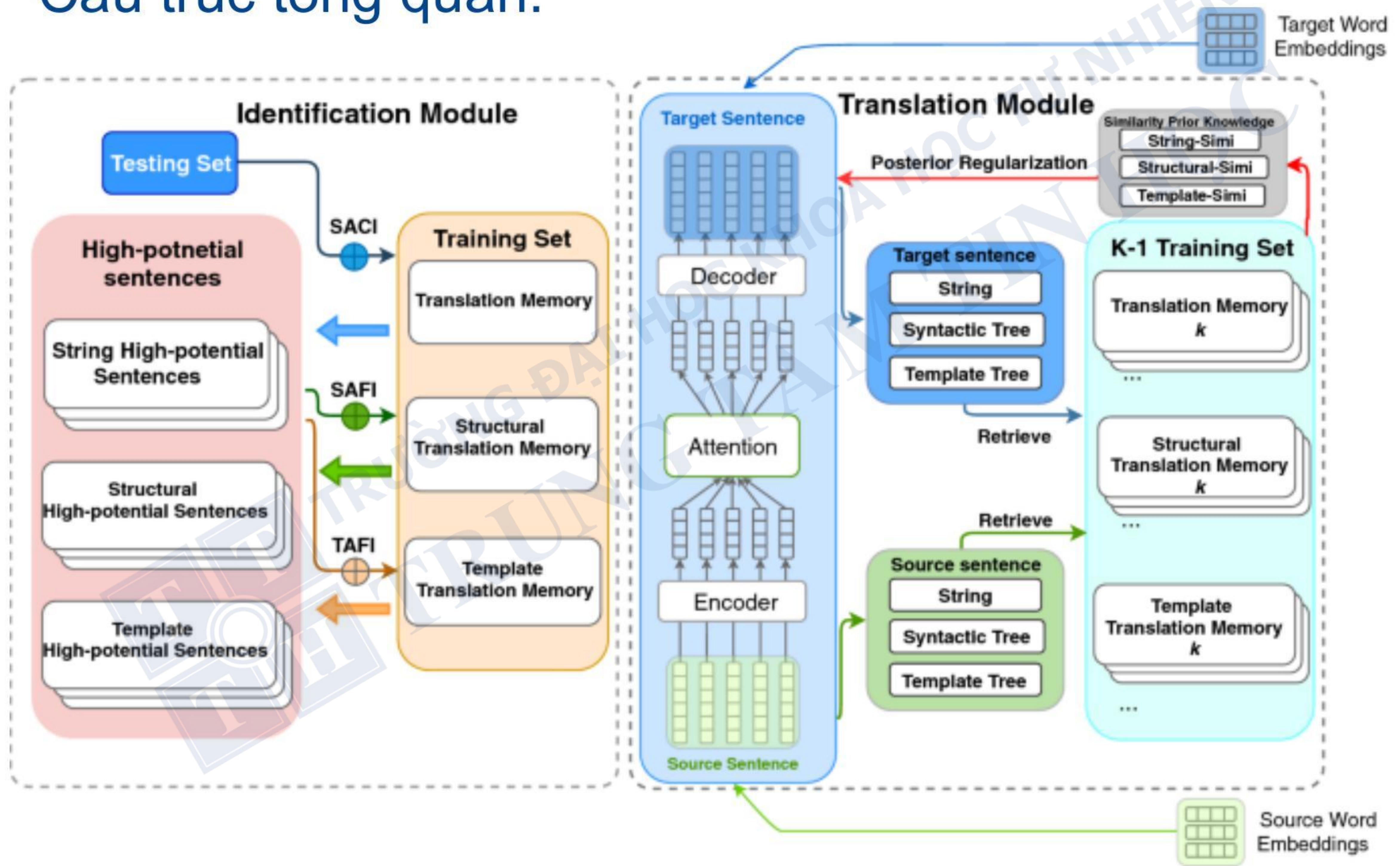
1. Học và hiểu các mối quan hệ phức tạp giữa các từ và câu.
2. Dịch ngôn ngữ với chất lượng cao và tự nhiên hơn so với các phương pháp truyền thống.
3. Tự động học từ dữ liệu và cải thiện hiệu suất theo thời gian.

→ Ứng dụng phổ biến trong dịch ngôn ngữ tự động.



Neural Machine Translation

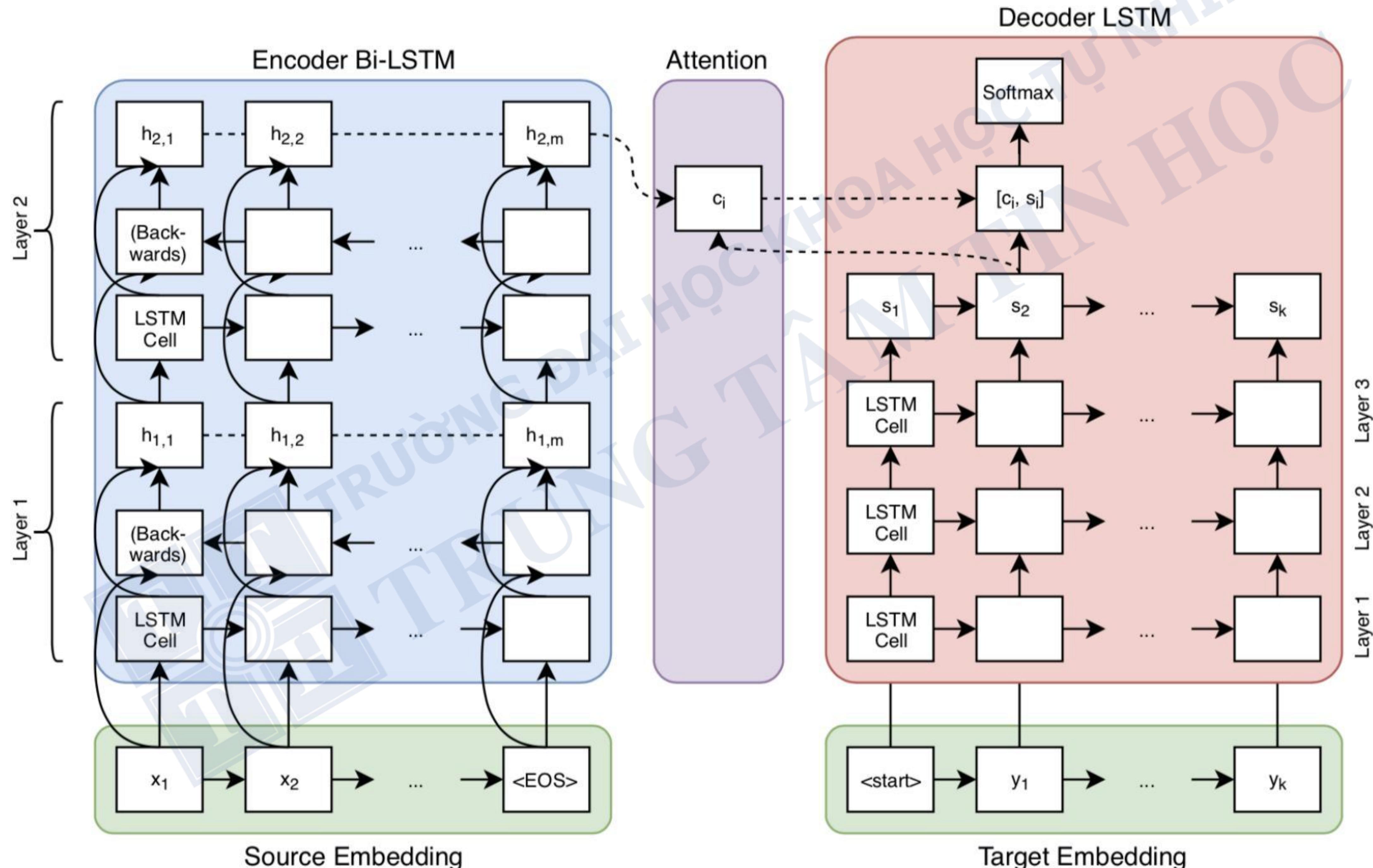
Cấu trúc tổng quan:





Neural Machine Translation

Encoder – Decoder:

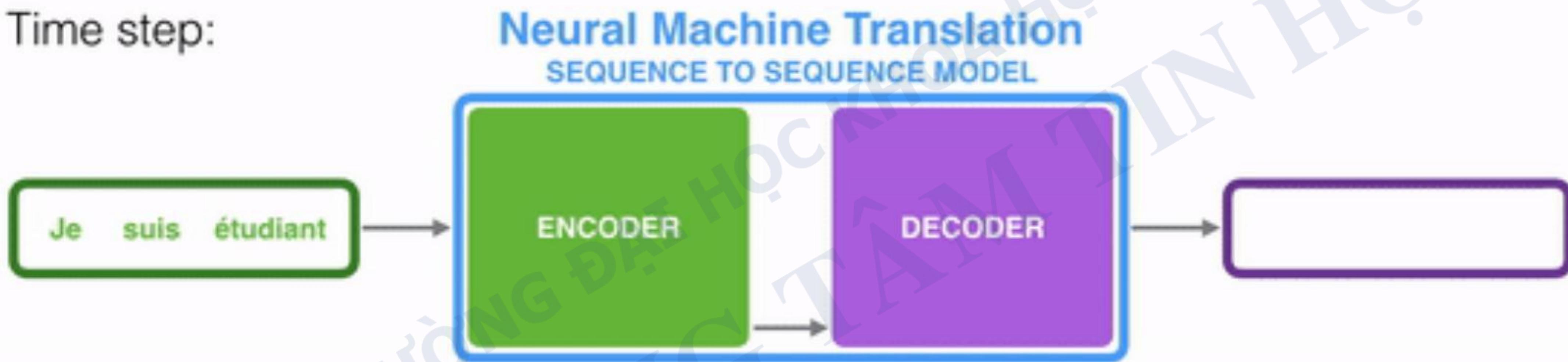




Neural Machine Translation

Ví dụ trực quan:

Time step:





Neural Machine Translation

Encoder truyền một lúc toàn bộ các hidden states đến decoder.

Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL WITH ATTENTION



Decoder thông qua attention tập trung vào các từ quan trọng hơn.

Neural Machine Translation



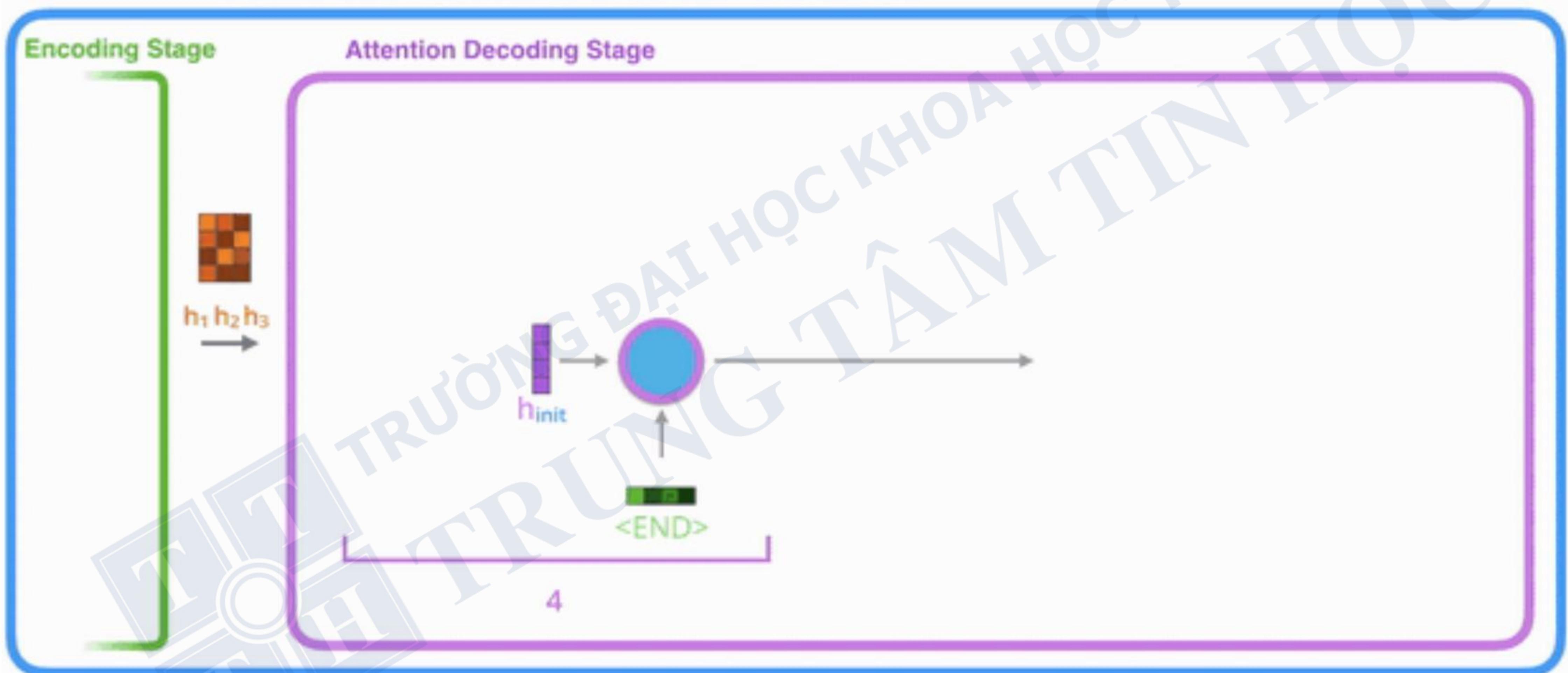
Attention at time step 4



Neural Machine Translation



Neural Machine Translation SEQUENCE TO SEQUENCE MODEL WITH ATTENTION



Code Demo



DEMO



Q&A

