****

**Đặng Nhật Linh ()**

**Phạm Quốc Hoàng (2221050234)(C)**

**ĐỀ TÀI**

**Dịch máy tự động (Machine Translation)**

BÀI TẬP LỚN MÔN: XỬ LÝ NGÔN NGỮ TỰ NHIÊN

**Hà Nội, Năm 2024**

MỤC LỤC

# Giới thiệu

* 1. **Định nghĩa dịch máy**
* **Dịch máy** (hay còn được gọi là dịch tự động) là quá trình sử dụng các phần mềm dịch thuật được ứng dụng trí tuệ nhân tạo để thực hiện dịch một ngôn ngữ nào đó sang một hoặc nhiều ngôn ngữ khác một cách tự động.
* Ngôn ngữ của văn bản cần dịch được gọi là văn bản nguồn,ngôn ngữ của bản dịch được gọi là ngôn ngữ đích. Kết quả của văn bản đích có thể hiệu đính để trở thành bản dịch tốt, gần với ngôn ngữ tự nhiên của con người hơn.
  1. **Mục tiêu**
* Xây dựng một ứng dụng dịch máy sử dụng mô hình học sâu (Neural Machine Translation – NMT) để dịch tiếng anh sang tiếng việt.
* Bảo toàn ngữ cảnh liên câu và độ chính xác trong việc dịch.
  1. **Phạm vi**
* Dịch văn bản tiếng Anh.
* Dịch từ tiếng Anh sang tiếng Việt.
* Giao diện đơn giản.

1. **Mô tả vấn đề**
   1. **Nhu cầu người dùng**

* Việc dịch ngôn ngữ là cần thiết trong giao tiếp toàn cầu.
* Các công cụ dịch truyền thống thường không hiểu ngữ cảnh.
* Cần một hệ thống có khả năng dịch tự nhiên, mượt mà, hiểu ngữ cảnh câu.

* 1. **Thách thức**
* Ngữ pháp và cấu trúc khác nhau giữa các ngôn ngữ.
* Từ đa nghĩa, ngữ cảnh phức tạp.
* Dữ liệu huấn luyện lớn, yêu cầu tài nguyên cao.
  1. **Giải pháp**
* Sử dụng mô hình **Transformer** hoặc **seq2seq (LSTM/GRU)** với cơ chế Attention.
* Huấn luyện trên tập dữ liệu song ngữ.
* Tiền xử lý: chuẩn hóa, tokenization, embedding.
* Giao diện demo đơn giản (web app hoặc notebook).

1. **Mô hình dịch máy (Neural Machine Translation)**

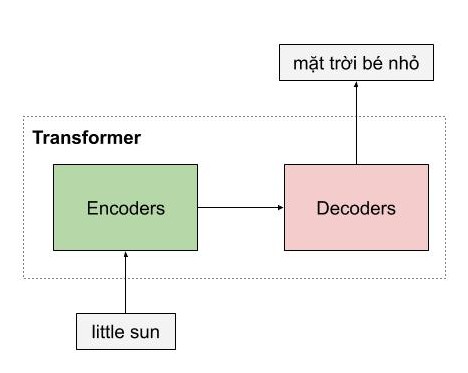
**3.1 Neural Machine Translation – NMT**

* **Neural** **Machine Translation (NMT)** là một phương pháp dịch máy sử dụng mạng nơ-ron sâu để tự động dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác.
* **Mục tiêu**: Dịch câu từ ngôn ngữ nguồn (source) sang ngôn ngữ đích (target).
* Hệ thống NMT học cách ánh xạ toàn bộ câu đầu vào sang toàn bộ câu đầu ra bằng cách huấn luyện trên tập dữ liệu song ngữ lớn.

**3.2 Kiến trúc cơ bản**

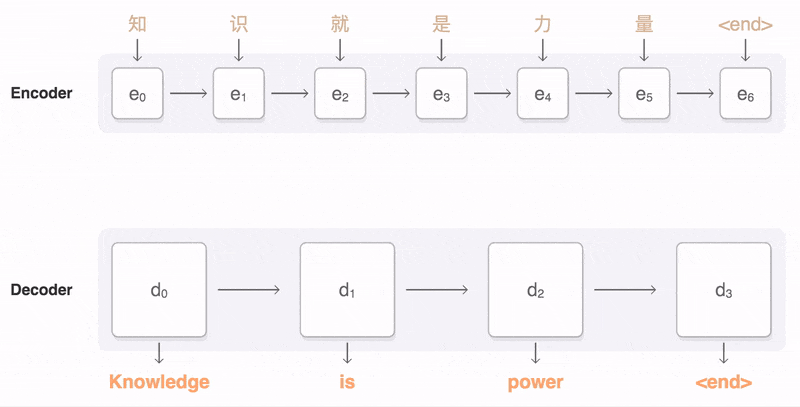
**a. Encoder-Decoder**

* **Encoder :** Mã hóa câu nguồn thành một vector ngữ nghĩa (context vector).
* **Decoder :** Giải mã vector đó để sinh ra câu đích.



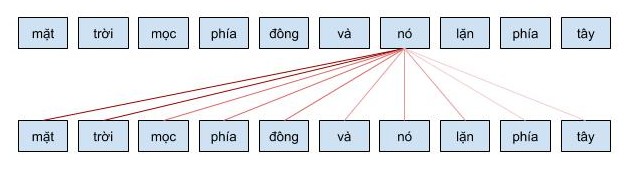
**b. Attention Mechanism**

* Cho phép mô hình tập trung vào các phần cụ thể của câu nguồn trong quá trình dịch → cải thiện độ chính xác.



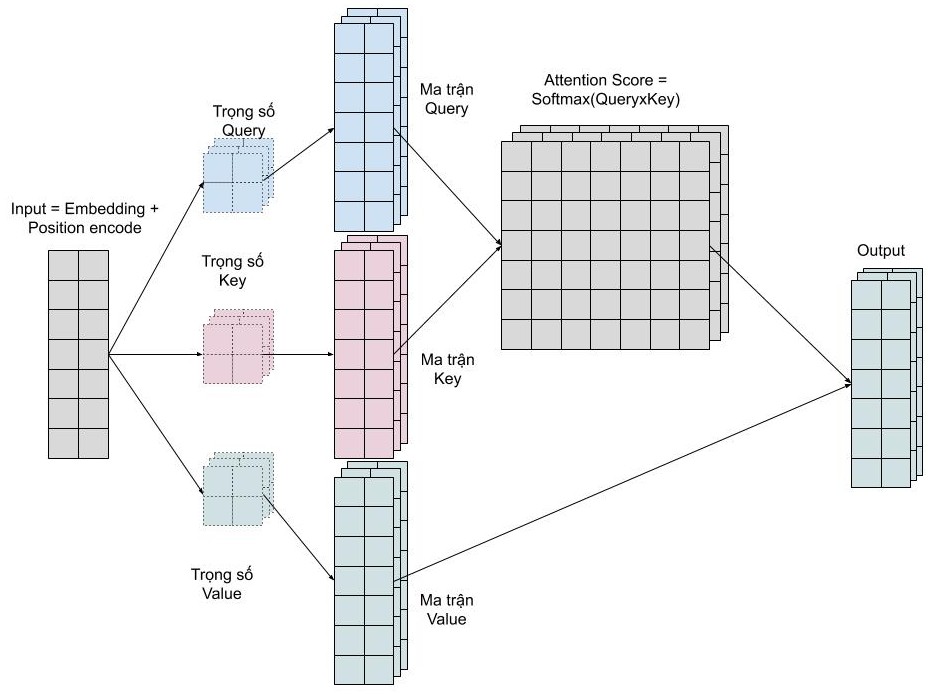
**c. Transformer (hiện đại nhất)**

* Kiến trúc không dùng RNN mà dùng self-attention để xử lý chuỗi song song hiệu quả hơn.
* Self Attention cho phép mô hình khi mã hóa một từ có thể sử dụng thông tin của những từ liên quan tới nó. Ví dụ khi từ **nó** được mã hóa, nó sẽ chú ý vào các từ liên quan như là **mặt trời**.

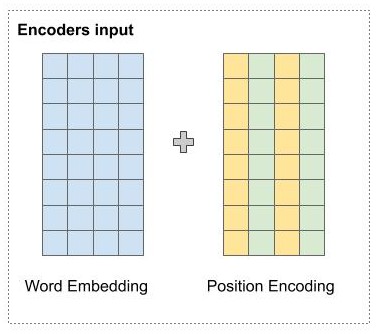


* Cấu trúc của Transformer:

+ Multi-head attention : Mô hình có thể học nhiều kiểu mối quan hệ giữa các từ với nhau. Với mỗi self-attention, chúng ta học được một kiểu pattern, do đó để có thể mở rộng khả năng này, chúng ta đơn giản là thêm nhiều self-attention.



+ Positional encoding :  Dùng để đưa thông tin về vị trí của các từ vào mô hình transformer, giúp mô hình nhận biết được thứ tự từ ngữ mà không cần xử lý tuần tự.



+ Feed-forward layers : Sau lớp Attention, mỗi vị trí trong câu được xử lý qua một mạng neural đơn giản để tăng khả năng biểu diễn.

+ Residual connections & Layer normalization : Hai kỹ thuật giúp cho mô hình huấn luyện nhanh hội tụ hơn và tránh mất mát thông tin trong quá trình huấn luyện mô hình.

**3.3 Ưu – Nhược điểm**

|  |  |
| --- | --- |
| **Ưu điểm** | **Nhược điểm** |
| **Chất lượng bản dịch cao** Dịch trôi chảy, tự nhiên, sát nghĩa hơn SMT | **Cần nhiều dữ liệu song ngữ** Hiệu quả giảm rõ nếu dữ liệu ít |
| **Hiểu ngữ cảnh toàn câu** Không dịch từng cụm rời rạc như SMT | **Tốn tài nguyên** Yêu cầu GPU mạnh, thời gian huấn luyện lâu |
| **Tự học biểu diễn ngữ nghĩa** Không cần thiết kế đặc trưng thủ công | **Là mô hình "hộp đen"** Khó giải thích hoặc điều chỉnh quy tắc dịch |
| **Dễ mở rộng với dữ liệu lớn** Không cần thay đổi mô hình khi có thêm dữ liệu | **Lỗi dịch "nghe có vẻ đúng" nhưng sai ngữ nghĩa** Khó phát hiện, nguy hiểm hơn |
| **Hỗ trợ học liên ngôn ngữ** Có thể huấn luyện mô hình đa ngôn ngữ | **Không phù hợp với ngôn ngữ ít tài nguyên** |

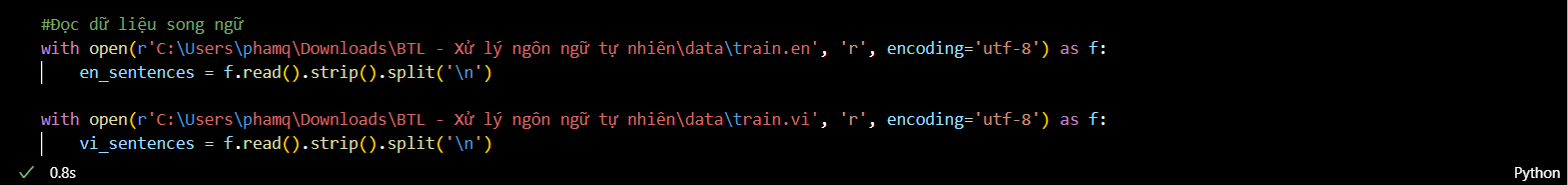
1. **Quy trình huấn luyện NMT**

**4.1 Chuẩn bị dữ liệu**

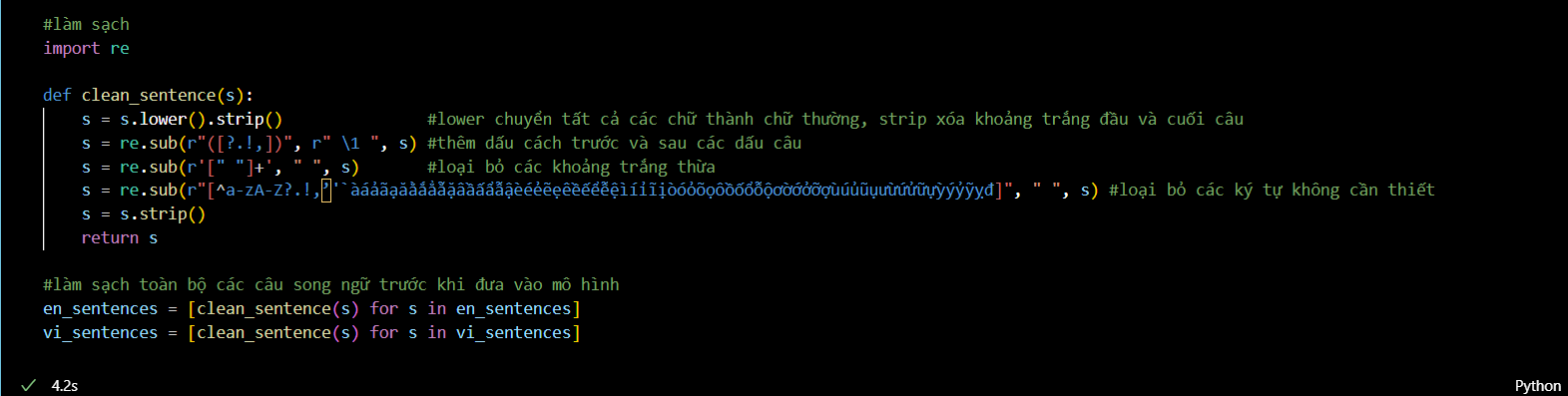
* Chuẩn bị tập văn bản song ngữ ( Anh – Việt ) : bộ dữ liệu song ngữ được thu thập trên TED bao gồm hơn 600k câu song ngữ anh-việt

**4.2 Tiền xử lý dữ liệu**

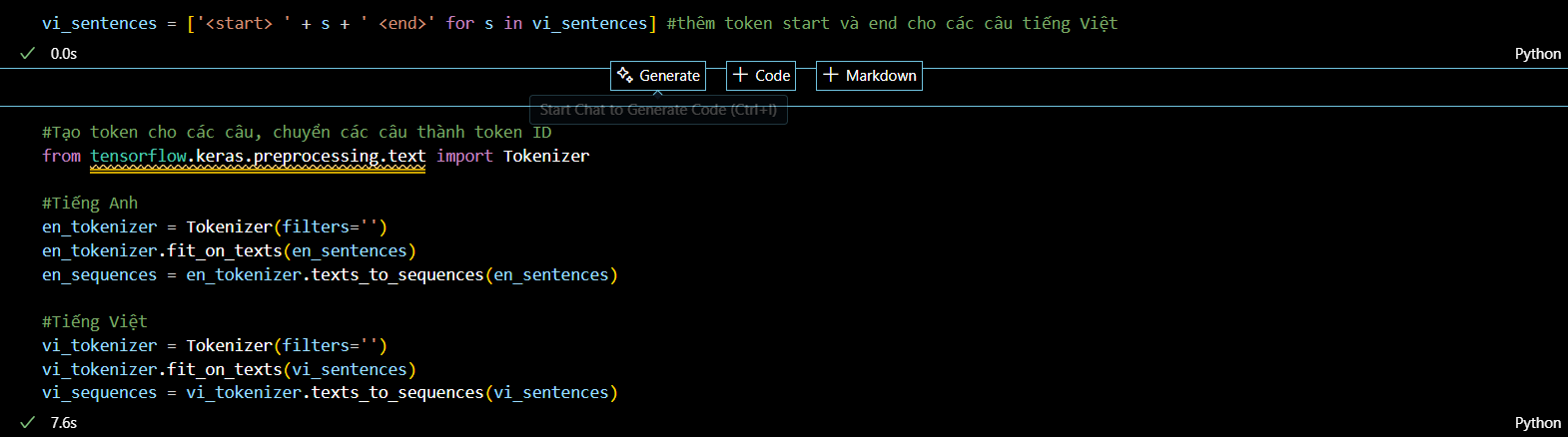
* Chuẩn bị và làm sạch dữ liệu thô trước khi sử dụng nó cho các bước phân tích, mô hình hóa hoặc học máy.
* Mục đích của tiền xử lý là biến dữ liệu đầu vào trở nên “sạch”, có cấu trúc rõ ràng và phù hợp hơn để các thuật toán có thể hoạt động hiệu quả, chính xác hơn.
* Bước 1: Đọc dữ liệu song ngữ đã chuẩn bị



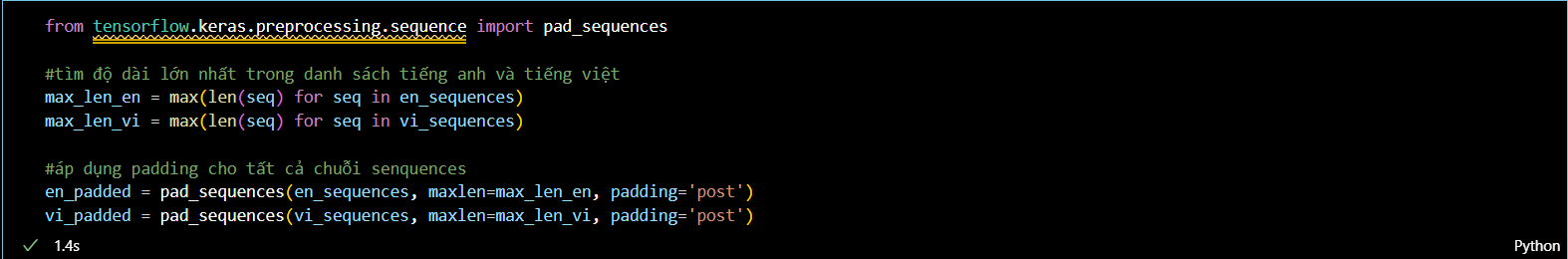
* Bước 2: Làm sạch



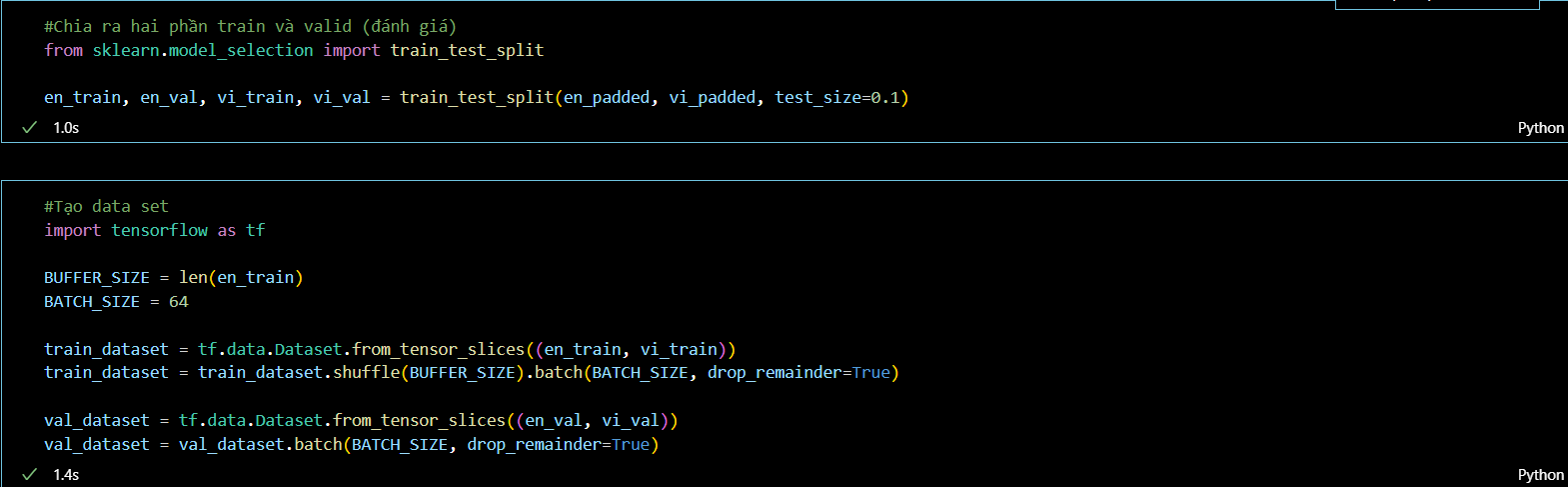
* Bước 3: Tokenizer



* Bước 4: Padding

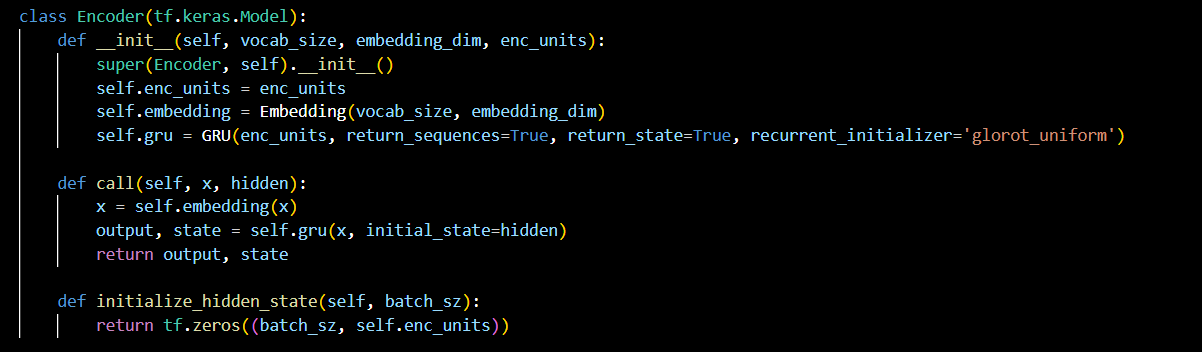


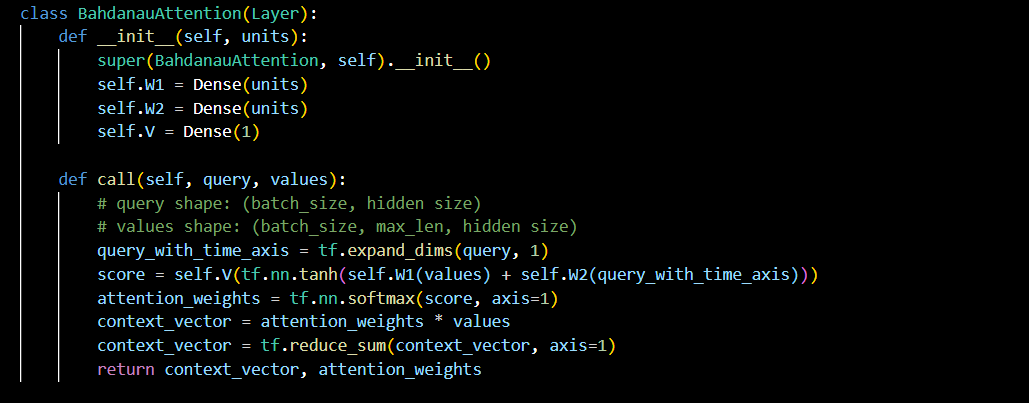
* Bước 5: Tạo Dataset để huấn luyện mô hình

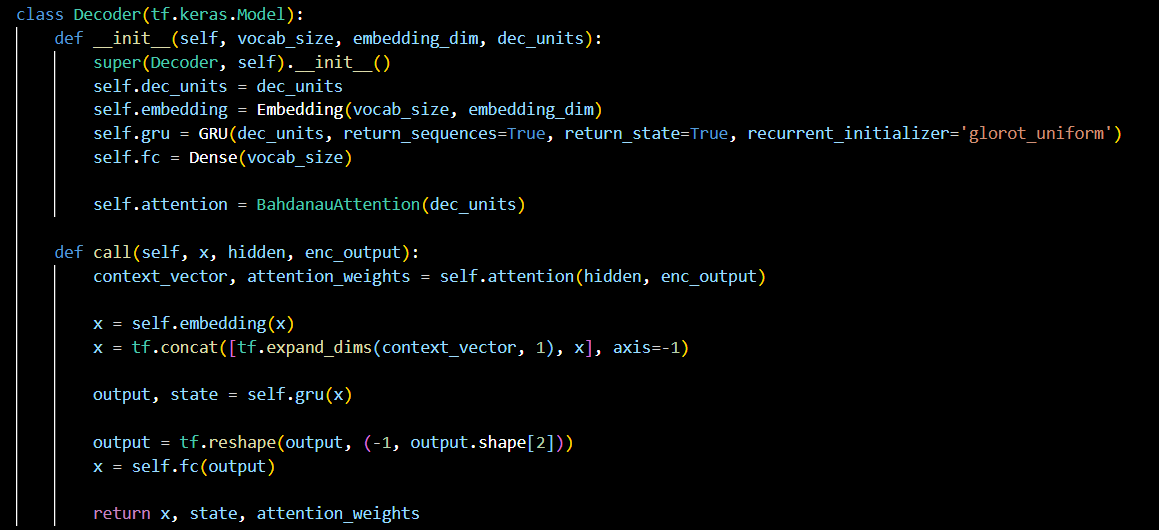


**4.3 Xây dựng mô hình**

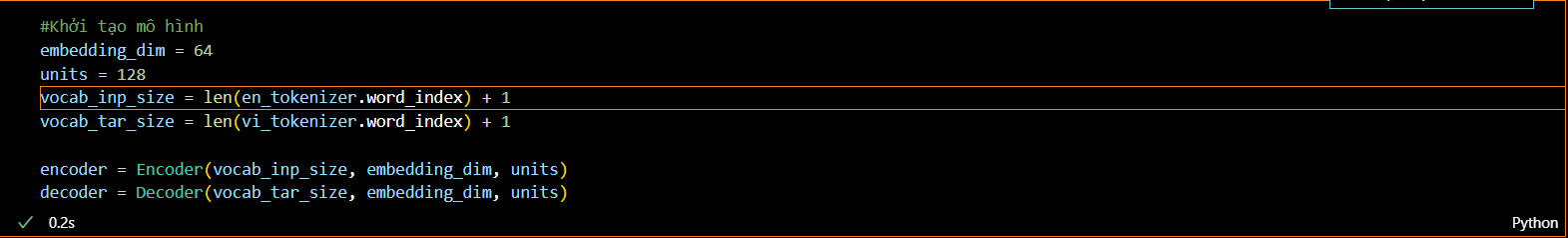
* Sử dụng Encoder, Attention, Decoder ( theo mô hình Seq2Seq )







* Sau khi đã định nghĩa thì sẽ bắt đầu khởi tạo mô hình:



**4.4 Huấn luyện mô hình**

**4.5 Đánh giá mô hình**

1. **Demo**

**Link github:**

1. **Tài liệu tham khảo  
   1.** [**https://eleks.com/research/neural-machine-translation-attention-mechanism/**](https://eleks.com/research/neural-machine-translation-attention-mechanism/)

**2.** [**https://pbcquoc.github.io/transformer/**](https://pbcquoc.github.io/transformer/)

1. **Kết luận**
2. **Phân chia công việc**

****

**BÀI TẬP LỚN**

**MÔN:**

**ĐỀ TÀI**

**SINH VIÊN THỰC HIỆN Cán Bộ giảng dạy**

Đặng Nhật Linh ()

Phạm Quốc Hoàng (2221050234)(C)

**Hà Nội, Năm 2024**