Deep Learning in Business
“Attention is all you need”

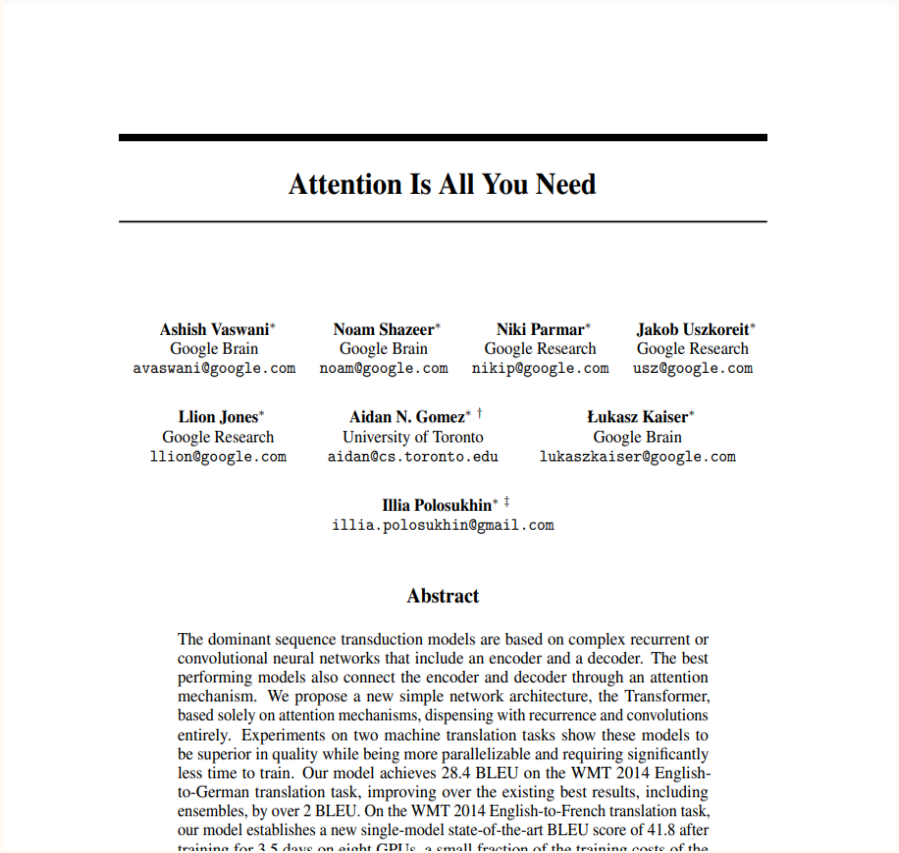
Giáo viên: GS.TS ĐỖ PHÚC
Trợ giảng: NCS PHAN HỒNG TRUNG
Thành viên: Đặng Thị Huệ
Vũ Thị Minh Phương

Nội dung trình bày

1. Giới thiệu
2. Tại sao lại là Transformer ?
3. Kiến trúc mô hình Transformer
4. Thử nghiệm
5. Tài liệu tham khảo

1. Giới thiệu

- Năm 2017, Google công bố bài báo "Attention Is All You Need" thông tin về Transformer như tạo ra bước ngoặt mới trong lĩnh vực xử lý ngôn ngữ tự nhiên



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

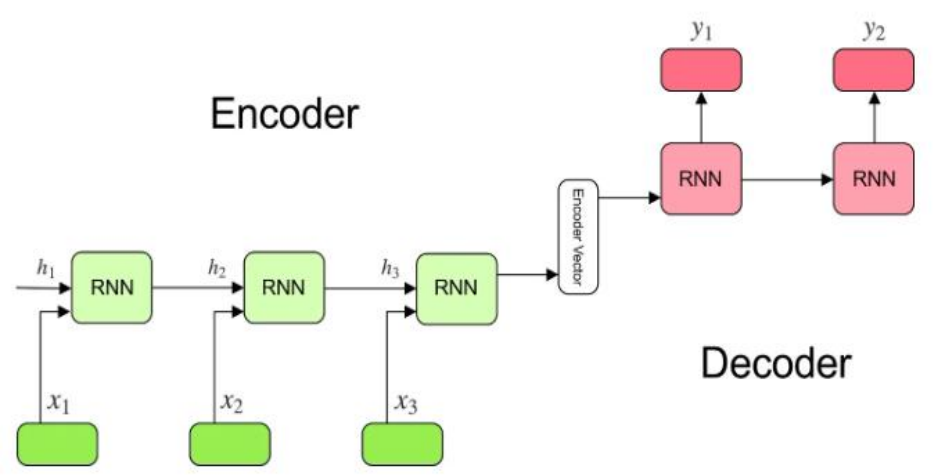
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the

2. Tại sao lại là Transformer ?



Mô hình RNN trong dịch máy

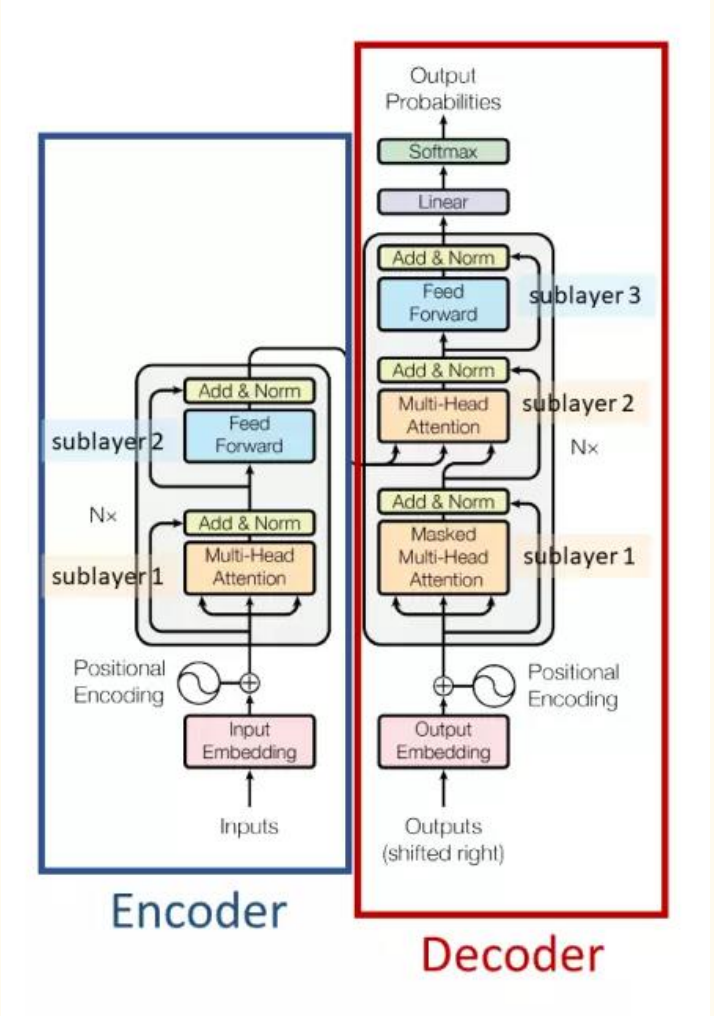
- CNN: Kiến trúc này yêu cầu đầu vào có kích thước cố định và không thể xử lý các đầu vào có độ dài khác nhau, gây khó khăn với các bài toán xử lý ngôn ngữ tự nhiên với độ dài của câu có thể khác nhau

- RNN: Do phải xử lý câu đầu vào một cách tuần tự nên nhược điểm của mô hình này là tốc độ xử lý chậm và hạn chế trong việc biểu diễn sự phụ thuộc xa giữa các từ trong một câu. Vấn đề vanishing gradient cũng gây nhiều khó khăn gây ảnh hưởng đến khả năng học tập của mô hình và làm giảm hiệu suất của nó.

➔ **Mô hình Transformer với cơ chế self-attention xuất hiện giải quyết các khó khăn trên**

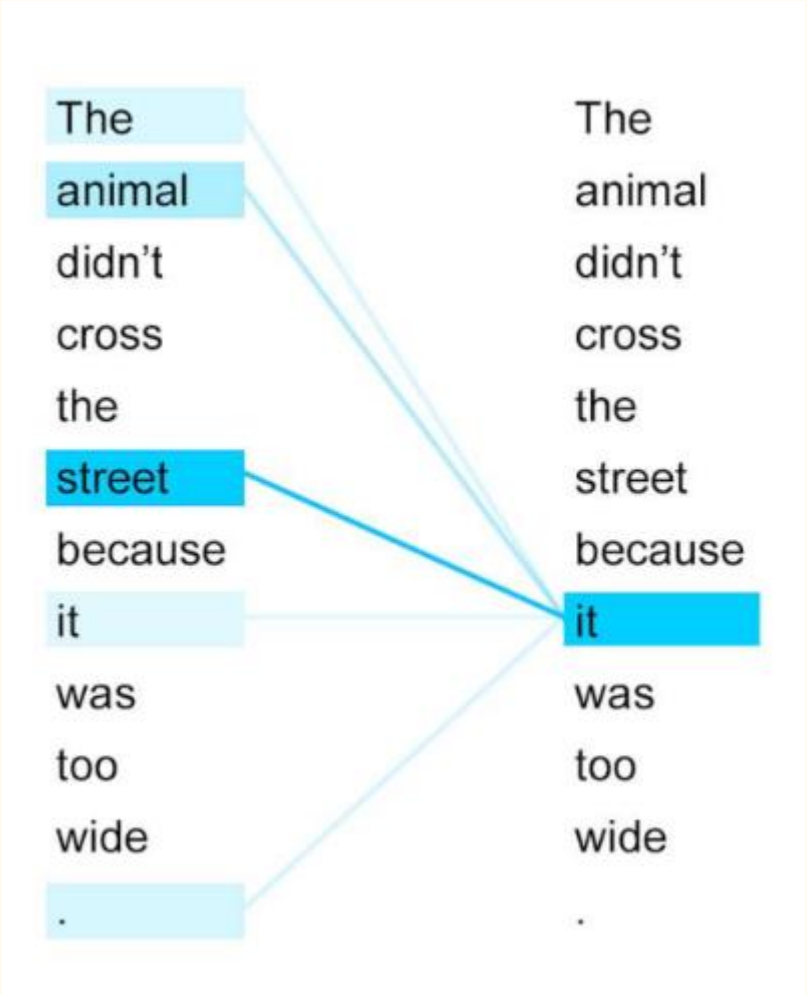
3. Mô hình Transformer

- Encoder: 6 lớp giống nhau, mỗi lớp gồm 2 lớp con
- Decoder: 6 lớp giống nhau,mỗi lớp gồm 3 lớp con



3.1. Seft-attention

- Đây là cốt lõi và là linh hồn của Transformer
- Cơ chế tạo ra quan hệ các từ trong câu
- Khi được mã hóa (encoder) nó sẽ mang nhiều thông tin của các từ liên quan

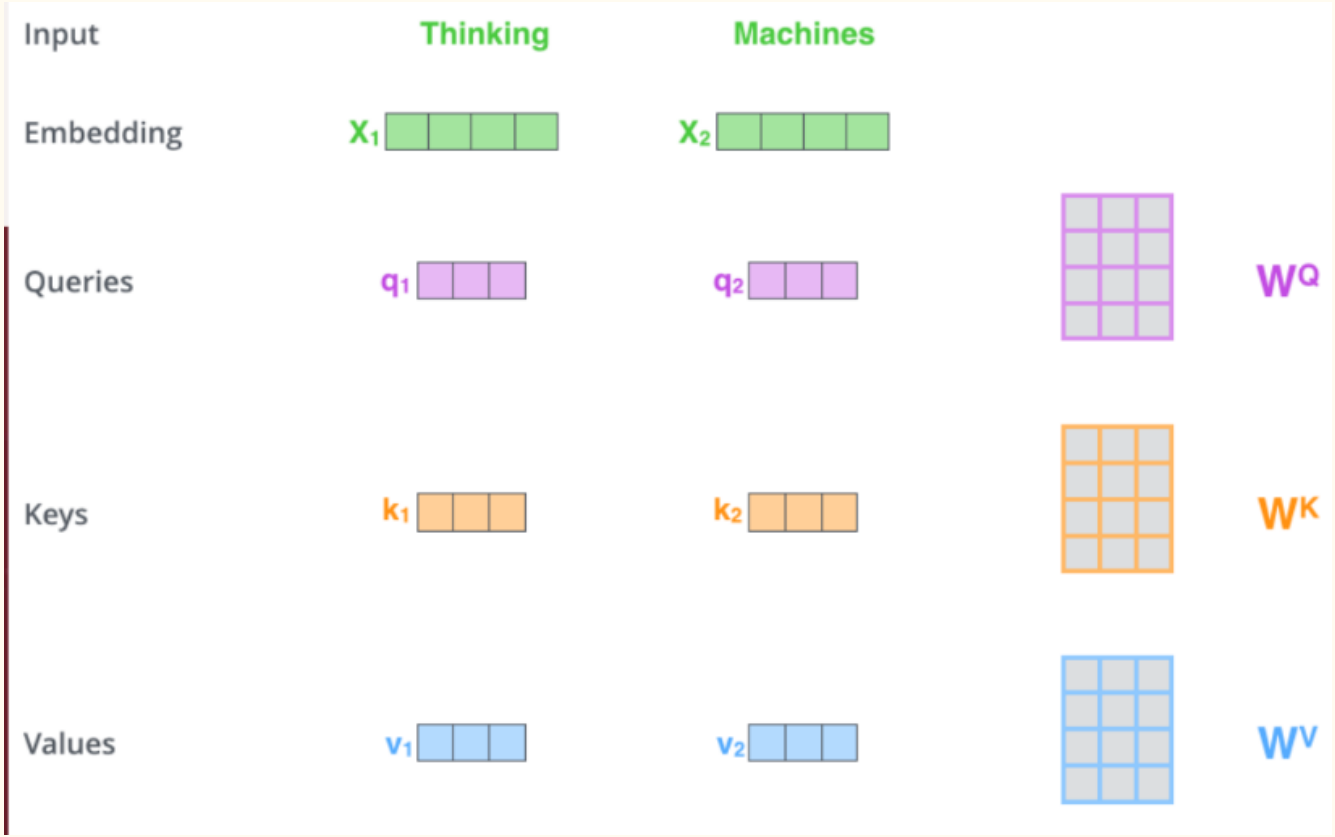


3.1. Seft-attention

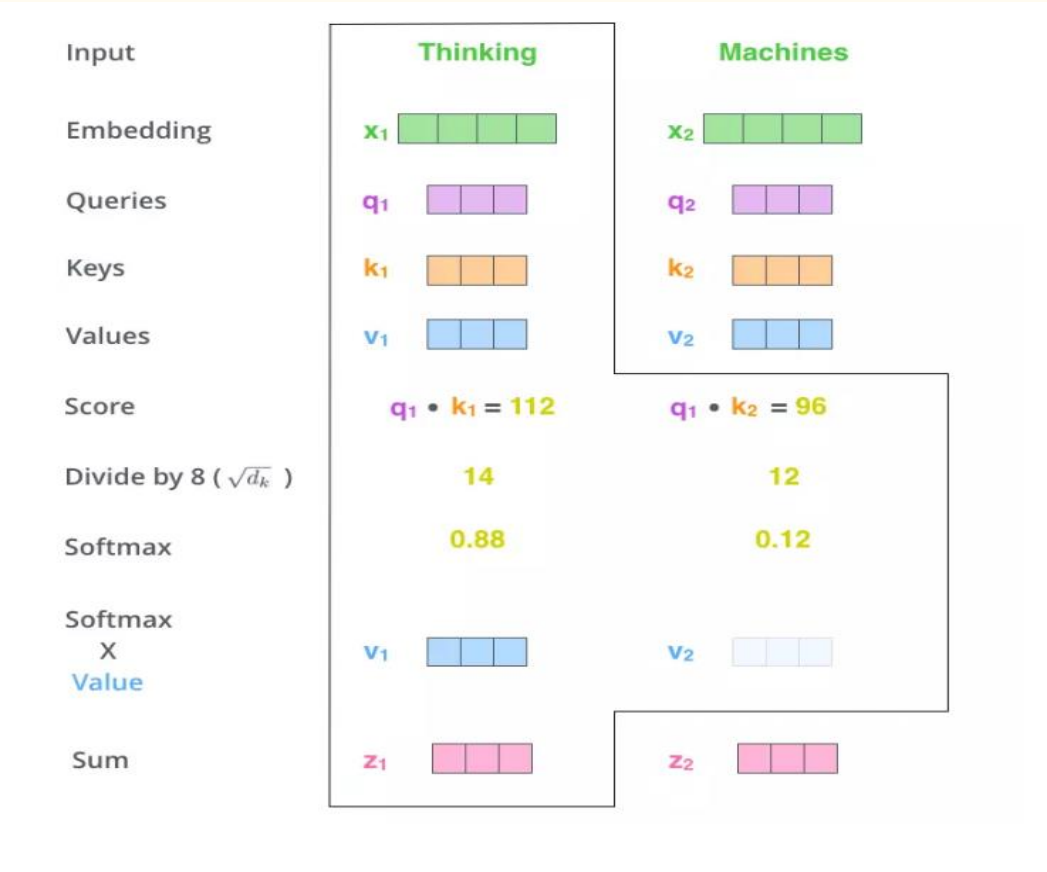
Đầu vào của self-attention là 3 vector query, key, value. Các vector này được tạo ra bằng cách nhân ma trận biểu diễn các từ đầu vào với ma trận học tương ứng.

- query vector là vector dùng để chứa thông tin của từ được tìm kiếm, so sánh.
- key vector là vector dùng để biểu diễn thông tin các từ được so sánh với từ cần tìm kiếm ở trên.
- value vector là vector biểu diễn nội dung, ý nghĩa của các từ

3.1. Seft-attention



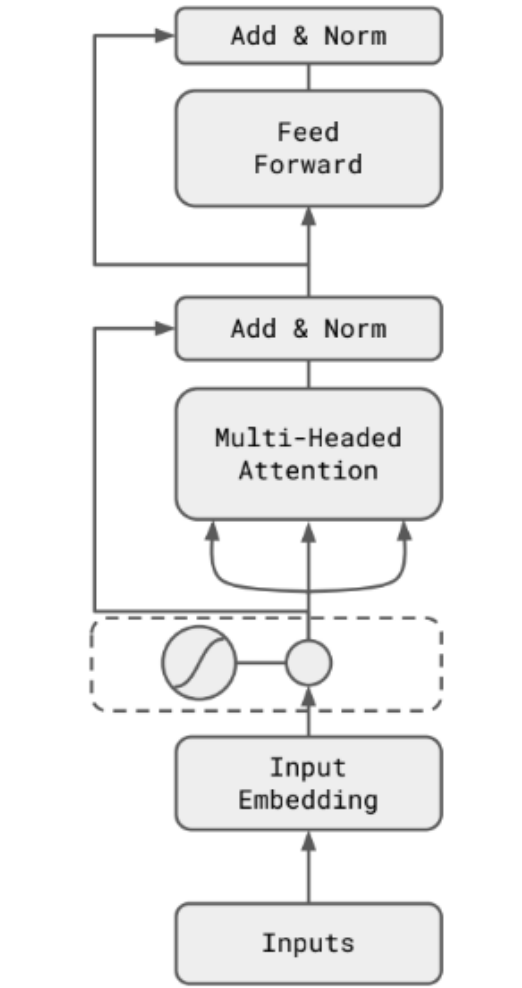
3.1. Seft-attention



Quá trình tính toán self-attention

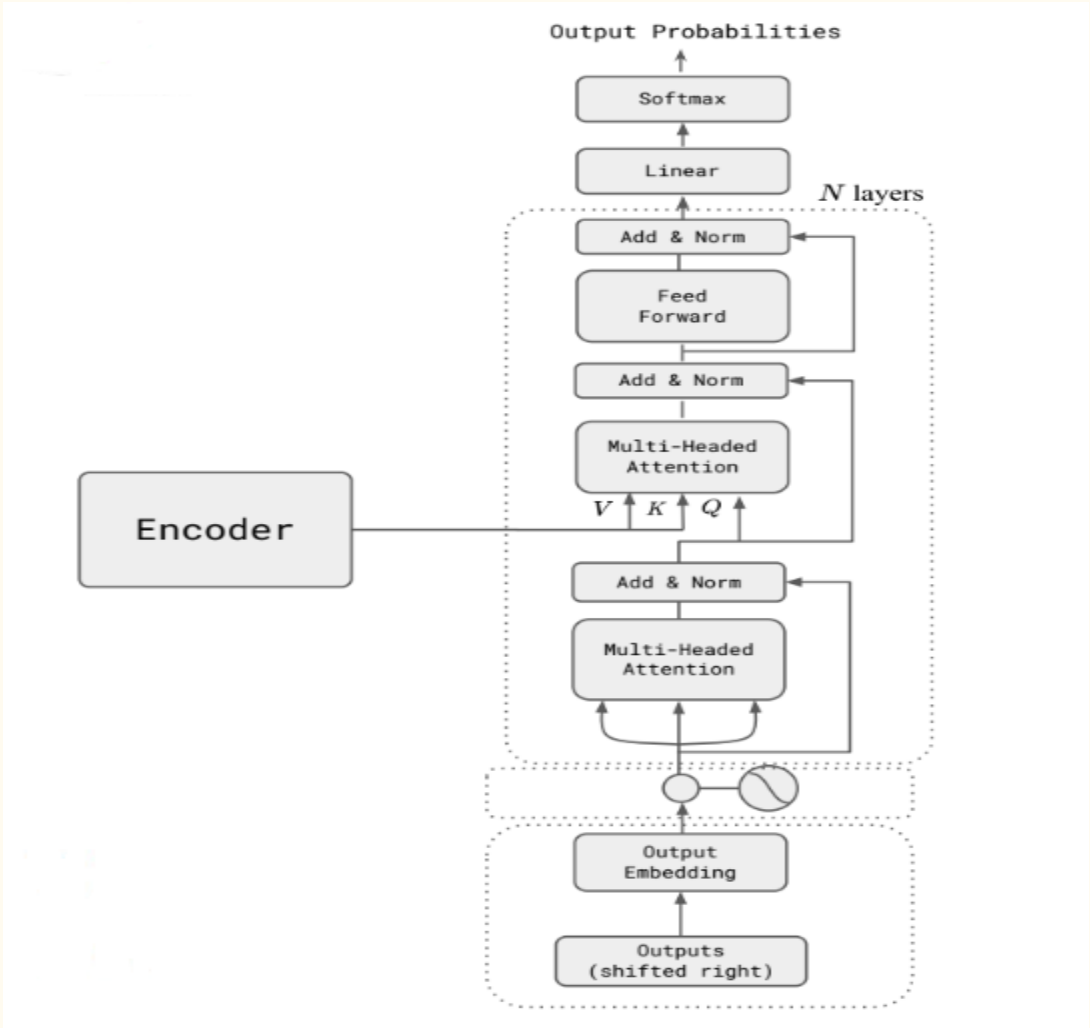
3.2. Encoder

- Input Embedding: Các câu đầu vào sẽ được mã hóa thành các vector bằng việc sử dụng Word Embedding
- Positional Encoding: cho biết thêm thông tin về vị trí của một từ
- Multi-Head attention: nhiều seft-attention
- Feed Forward;



3.3. Decoder

- Output Embedding
- Masked multi-head attention
- Multi-Head attention
- Position-wise fully connected feed forward neural network



4. Thử nghiệm

Bộ dữ liệu Multik30k của thư viện torchtext, bộ dữ liệu gồm:

- Tập huấn luyện (Training set): 29000 câu văn bản tiếng Anh và tiếng Đức
- Tập đánh giá (Valid): bao gồm 1014 cặp câu văn bản tiếng Anh và tiếng Đức
- Tập kiểm tra (Test set): bao gồm khoảng 1000 cặp câu văn bản tiếng Anh và tiếng Đức

Kết quả, độ chính xác mô hình

```
bleu_score = calculate_bleu(test_data, SRC, TRG, model, device)

print(f'BLEU score = {bleu_score*100:.2f}')
```

```
BLEU score = 36.11
```

➔ BLEU score là một phương pháp đánh giá tự động tiêu chuẩn và phổ biến trong lĩnh vực dịch máy, giúp đo lường chất lượng dịch và so sánh các hệ thống dịch máy khác nhau.

5. Kết luận

- Kiến trúc Transformer cho phép thực hiện các phép tính song song nên giảm đáng kể thời gian huấn luyện, tận dụng được sức mạnh tính toán của multi-GPU.
- Transformer ra đời kế thừa ý tưởng từ self attention từ LSTM, loại bỏ hoàn toàn tính tuần tự phụ thuộc hoàn toàn vào cơ chế attention để tính toán ra được mối tương quan giữa input và output

Tài liệu tham khảo

[1] Bui, M. Q. (2021, March 29). *Tản mạn về Self Attention*. Viblo. Retrieved May 12, 2023, from https://viblo.asia/p/tan-man-ve-self-attention-07LKXoq85V4

[2] Nguyen, A. V. (2020, May 1). *Transformers - "Người máy biến hình" biến đổi thế giới NLP*. Viblo. Retrieved May 12, 2023, from https://viblo.asia/p/transformers-nguoi-may-bien-hinh-bien-doi-the-gioi-nlp-924lJPOXKPM

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N. Gomez, A., Kaiser, L., & Polosukhin, I. (2017, June 12). *[1706.03762] Attention Is All You Need*. arXiv. Retrieved May 12, 2023, from https://arxiv.org/abs/1706.03762

[4] Link code mô hình Transformer (chỉnh sửa bổ sung thư viện) Google Colab do nhóm báo cáo đã chạy tại đây

TRÂN TRỌNG
CẢM ƠN QÚY THẦY