AI VIET NAM – COURSE 2022

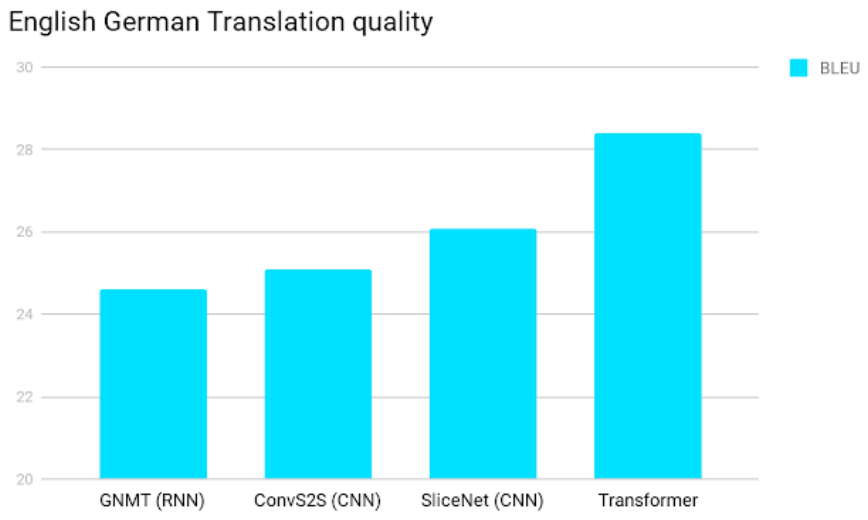
Attention Is All You Need

Ngày 30 tháng 11 năm 2022

Date of publication:	12/06/2017
Authors:	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin
Sources:	Attention Is All You Need
Data sources:	Transformer
Keywords:	Attention, Transformer, Seq2Seq
Summary by:	Quoc Viet

1. **Mục đích:**

Trên thang điểm đánh giá nhiệm vụ phiên dịch từ tiếng Anh sang tiếng Đức, mô hình Transformer [3] vượt qua tất cả các mô hình mạng hồi quy và mạng tích chập, như ở hình 1. Bên cạnh việc cải thiện chất lượng dịch thuật, mô hình Transformer sử dụng ít tính toán hơn để huấn luyện và phù hợp hơn với các phần cứng máy học hiện đại. Điều này cho phép việc huấn luyện được hoàn thành tỏng khoản thời gian ngắn hơn.



Hình 1: Điểm BLEU (cao hơn là tốt hơn) của từng mô hình trên thang đánh giá dịch thuật từ tiếng Anh sang tiếng Đức. (*Nguồn: Google AI Blog*)

2. **Đóng góp:**

Transformer là mô hình đầu tiên tính toán các biểu diễn (representations) của đầu ra và đầu vào chỉ sử dụng cơ chế self-attention và loại bỏ mạng hồi quy thần kinh (Recurrent Neural Netwwork) và tích chập (Convolution Neural Network). Phần mã hoá đầu ra (Encoder) của mộ hình Transformer đọc một chuỗi các từ cùng lúc, điều này trái ngược với các mô hình có hướng (directional models), khi mô hình có hướng đọc các văn bản một cách tuần tự (từ trái sang phải hoặc từ phải sang trái). Kết quả là, mô hình Transformer là một mô hình song hướng (bidirectional), đặc trưng này giúp mô hình học được nội dung của các từ dựa trên những từ xung quanh của nó.

3. **Phương pháp luận:**

- **Attention**

Cơ chế Attention [1] nhận các chuỗi đầu vào và xác định xem các yếu tố nào trong chuỗi có liên quan đến từng giai đoạn. Nói cách khác, Attention cho phép mô hình tập trung vào các từ liên quan đến chuỗi dữ liệu đầu vào khi cần. So với mô hình chuỗi sang chuỗi (sequence-to-sequence) truyền thống [2], trong mô hình Attention, phần mã hoá đầu vào (Encoder) truyền đi nhiều dữ liệu hơn cho decoder, thay vì chỉ truyền dữ liệu ở trạng thái cuối cùng của gia đoạn mã hoá đầu vào, encoder truyền toàn bộ dữ liệu từ tất cả các trạng thái ẩn (hidden states) đến decoder.

- **Mô hình Attention sử dụng tích vô hướng (Scaled Dot-Product Attention)**

Mô hình Attention sử dụng tích vô hướng là một mô hình có cấu trúc như ở Hình 2, kết quả của các tích vô hướng sẽ được chia xuống với đại lượng $\sqrt{d_k}$. Mô hình Attention sử dụng tích vô hướng có thể được mô tả bởi phương trình sau:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

trong đó:

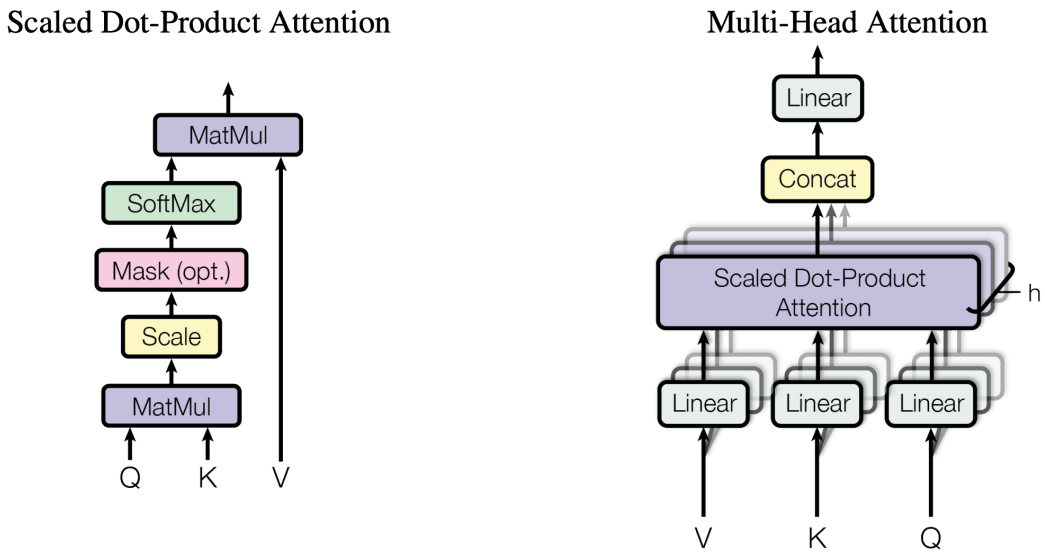
- Q là ma trận chứa các truy vấn (query) là các biểu diễn dạng vector của một từ trong chuỗi
- K là tất cả các khoá (key), hay là biểu diễn vector của toàn bộ từ trong chuỗi
- V là các giá trị có biểu diễn vector của toàn bộ từ trong chuỗi

Trọng số $a = softmax(\frac{QK^T}{\sqrt{d_k}})$ thể hiện mức độ mà mỗi từ của chuỗi (được kí hiệu là Q) bị ảnh hưởng bởi các từ còn lại trong chuỗi (được kí hiệu bởi k).

- **Multi-Head Attention**

Multi-head Attention là một module thực thi các cơ chế Attention nhiều lần cùng lúc. Những đầu ra độc lập của cơ chế Attention sẽ được kết hợp lại và biến đổi tuyến tính đến số chiều nhất định. Multi-head Attention cho phép chú ý đến nhiều phần khác nhau trong chuỗi. Từng head của Attention có thể được xem như các tính toán tách biệt.
Cơ chế Multi-head Attention có thể được biểu diễn toán học như sau:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1, \ldots, \text{head}_h]\, \mathbf{W}_0,$$
$$\text{trong đó } head_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right)$$



Hình 2: Mô hình Attention sử dụng tích vô hướng (Scaled Dot-Product Attention) (bên trái). Multihead Attention gồm nhiều lớp Attention được chạy song song (bên phải)

- **Kiến trúc mô hình Transformer**

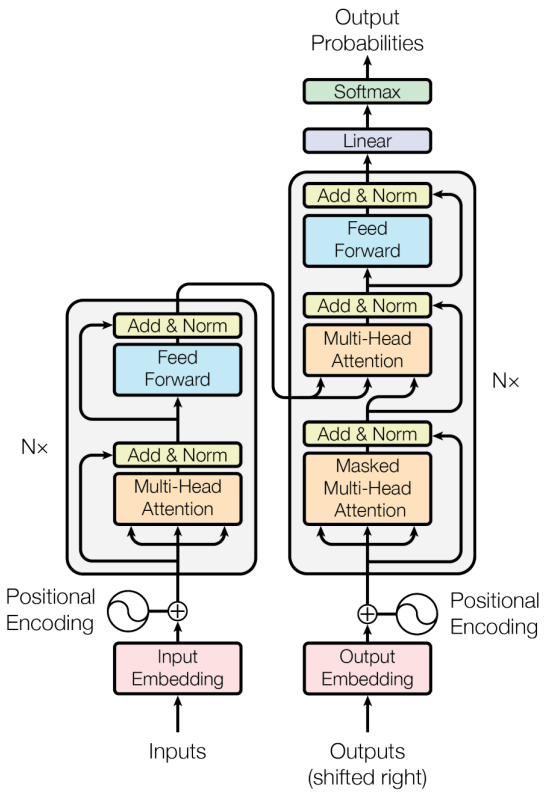
  Như ở Hình (3), kiến trúc mô hình Transformer theo một cấu trúc encoder-decoder nhưng không phụ thuộc vào hồi quy hay tích chấp để tạo ra đầu ra. Cả Encoder và Decoder đều chứa các module được xếp chồng lên nhau. Encoder nhận chuỗi đầu vào và ánh xạ lên một không gian nhiều chiều (thành một vector có n chiều). Vector này là đầu vào của decoder, decoder sau đó biến vector này thành chuỗi đầu ra tương ứng.

  Decoder sau đó xây dựng câu đầu ra theo từng từ một trong khi tra cứu các biểu diễn của encoder. Với từng từ, Transformer đầu tiên tạo ra một biểu diễn ban đầu. Sau đó, cơ chế self-attention sẽ tích luỹ thông tin của tất cả các từ khác, tạo ra một biểu diễn mới cho từng từ được dựa trên toàn bộ nội dung. Giai đoạn này được lặp lại cho tất cả các từ và được tính toán song song, kết quả là một chuỗi các biểu diễn mới.

  Đặc điểm chính của mô hình Transformer là từng từ sẽ có một đường đi riêng trong encoder. Ở layer self-attention, tồn tại các sự phụ thuộc lẫn nhau giữa các đường đi này. Tuy nhiên, vì ở lớp lan truyền thuận (feed-forward layer) không có những sự phụ thuộc này, các được đi khác nhau có thể được tính toán song song khi đi qua lớp lan truyền thuận.

- **Biểu diễn vị trí (Positional Embedding)**

  Để thêm thông tin vị trí cho mô hình, các tác giả đề xuất một phương pháp biểu diễn vị trí (Positional Embedding) với đầu ra là một vector d chiều của một vị trí cụ thể trong câu sử dụng hàm Sinusoid. Đặt t là vị trí mong muốn của câu đầu vào, $\overrightarrow{p_t} \in \mathbb{R}^d$ là mã hoá đầu vào tương ứng, và d là số chiều mã hoá đầu vào. Khi đó $f : \mathbb{N} \to \mathbb{R}^d$ là hàm với đầu ra là một vector $\overrightarrow{p_t}$ được định nghĩa bởi:



Hình 3: Kiến trúc mô hình Transformer

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k \cdot t), & \text{nếu } i = 2k \\ \cos(\omega_k \cdot t), & \text{nếu } i = 2k+1 \end{cases}$$

trong đó

$$\omega_k = \frac{1}{10000^{2k/d}}$$

4. **Kết quả:**

- Mô hình Transformer (lớn) ở Hình 4 vượt qua mô hình tốt nhất được báo cáo (bao gồm các phương pháp học củng cố (ensemble)), với 2.0 điểm BLEU hơn ở nhiệm vụ phiên dịch tiếng Anh sang tiếng Đức trên bộ dữ liệu WMT 2014, tạo nên điểm state-of-the-art (SOTA) mới là 28.4.

- Ở nhiệm vụ phiên dịch từ tiếng Anh sang tiếng Pháp trên bộ dữ liệu WMT 2014, mô hình Transformer đạt điểm BLEU là 41.0, vượt qua tất cả các mô hình đã từng được công bố trước đây, trong khi thời gian huấn luyện chỉ bằng 1/4 so với các mô hình SOTA trước.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Hình 4: Mô hình Transformer đạt được điểm BLEU cao hơn các mô hình SOTA trước trong bài test tiếng Anh sang tiếng Đức và tiếng Anh sang tiếng Pháp với chi phí huấn luyện nhỏ hơn đáng kể.

Tài liệu

[1] Luong, M.-T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[2] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks, 2014.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017.