BÁO CÁO BÀI TẬP TUẦN 7

Người làm: Trần Đức Thọ

Tuần này em đã đọc paper nhưng vẫn có nhiều chỗ khó hiểu, em đã tìm kiếm slide bài giảng Mô hình mạng Transformer trong môn học sâu và ứng dụng của trường, em hiểu và trình bày ý hiểu của mình theo slide bài giảng em tìm được

SELF-ATTENTION:

Tại sao phải sử dụng SELF-ATTENTION ?

Trong mạng RNN : RNNs xử lý tuần tự từ trái sang phải

- Vấn đề: RNNs cần O(độ dài dãy) bước để các cặp từ cách xa nhau tương tác.

Chart

Description automatically generated

-Trong RNN các từ các xa nhau thì ngữ nghĩa của chúng càng hạn hẹp.

🡪SELF-ATTENTION sinh ra để giải quyết phần nào vấn đề này

SELF-ATTENTION là gì ?

Là cơ chế chú ý làm việc với queries, keys và values.

Text

Description automatically generated

Trong self-attention, queries, keys và values sinh ra từ cùng một nguồn.

Text

Description automatically generated

Nhược điểm của SELF-ATTENTION: SELF-ATTENTION hoạt động dựa trên các tập hợp đầu vào, tức là nó không quan tâm tới thứ tự đầu vào.

A picture containing timeline

Description automatically generated

Cách giải quyết: ta cần phải mã hoá thứ tự của câu vào trong keys, queries, và values, sau đó chỉ cần cộng p(i) vào các đầu vào

Có 2 cách đó là :

-Biểu diễn thông tin qua hàm sin

+ Ưu điểm: Tính chu kỳ chỉ ra rằng có thể “vị trí tuyết đối” không quá quan trọng

Có thể ngoại suy cho những câu dài hơn vì chu kỳ lặp đi lặp lại!

+ Nhược điểm: Không học được; việc ngoại suy cũng không thật sự hiệu quả!

- Học Véc tơ biểu diễn vị trí

+ Ưu điểm: Linh hoạt: mỗi vị trí được học để khớp dữ liệu

+ Nhược điểm: Không thể ngoại suy cho các vị trí ngoài dải 1, … , T

Một nhược điểm trong phương pháp SELF-ATTENTION các lớp SELF-ATTENTION chỉ là các lớp tuyến tính xếp chồng lên nhau, dù có xếp bao nhiêu lớp tuyến tính thì vẫn chỉ là lớp tuyến tính

Cách giải quyết: chúng ta có thể thêm mạng feed-forward (ví dụ ReLU) đề hậu xử lí vector đầu ra ví dụ như hình sau:

Diagram

Description automatically generated

Với các trọng số w1, w2, wT khi chúng ta qua SELF-ATTENTION là các lớp tuyến tính chúng ta triển khai qua mạng feed-forward để hậu xử lí các vector đầu ra như sau :

Text

Description automatically generated

Một trong những thách thức tiếp theo của SELF-ATTENTION đó là “không nhìn vào tương lai” khi đoán 1 câu:

Cách giải quyết

Diagram, text

Description automatically generated

Với những từ ở tương lai chúng ta thiết lập thành âm vô cùng. Như vậy chúng ta có thể dự đoán các giá trị này, ví dụ:

Chart

Description automatically generated

ở dòng đầu tiên chúng ta chỉ có 1 từ START chúng ta cần dự đoán tất cả các từ còn lại dòng tiếp theo chúng ta có 2 từ Start và The và ohari dự đoán tất cả các từ còn lại.

Tóm lại các trở ngại và giải pháp để sử dụng phương pháp SELF-ATTENTION trong mạng nơ-ron:

A picture containing chart

Description automatically generated

Ở trên là em giải thích kĩ thêm về mô hình SELF-ATTENTION, là kiến thức bắt buộc phải hiểu trước khi đi vào mạng Transformer

Giới thiệu về mạng mạng transformer

* Mạng transformer là một mô hình học sâu được giới thiệu năm 2017, được dùng chủ yếu ở lĩnh vực xử lý ngôn ngữ tự nhiên (NLP)
* Mạng transformer được ứng dụng nhiều hơn mạng RNN vì nó cải thiện được những nhược điểm của mạng RNN .
* Mạng transformer gồm cả 2 khối bao gồm Encoder và Decoder :

Diagram

Description automatically generated

Kiến trúc của mạng tranformer như sau : gồm rất nhiều khối Encoder xếp chồng lên nhau

Diagram

Description automatically generated

Tìm hiểu vào Encoder trước chúng ta sẽ thấy Encoder gồm 2 khối chính Self-Attention và Feed Forward Neural Network

Graphical user interface

Description automatically generated with medium confidence

Các khối Encoder giống nhau nên ở đây em đi sâu vào 1 khối Encoder được phân tích như sau

Giả sử chúng ta có 1 Seq gồm 3 từ:

Diagram

Description automatically generated

Khác với mạng RNN , transformer đưa tất cả các từ vào Self-Attention thay vì chỉ đưa từng một. Self-Attention lúc này sẽ tạo mối quan hệ giữa các từ . Các từ trong vector x kết hợp với lớp Self-Attention để tạo thành vector z

Quá trình self attention được mô tả như sau

Diagram

Description automatically generated with medium confidence

Các ma trận query , key và value được khởi tạo ngẫu nhiên và được hiệu chỉnh trong quá trình train của mạng.

Tiếp theo em xây dựng công thức tính từ vector x sang vector z

Graphical user interface, application

Description automatically generated

Đầu tiên chúng ta lấy vector query nhân với vector k ta được 2 giá trị 112 và 96 , tiếp theo đem chia cho căn bậc 2 của giá trị số chiều vector k (trong bài là 64) =8 . được giá trị chúng ta đưa qua hàng Softmax ta thu được các giá trị.

Trong bài báo họ tiếp tục nhân với 8 bộ query , key và value

Chart, box and whisker chart

Description automatically generated

Họ nhân với từng bộ ma trận khác nhau để thu được các ngữ nghĩa khác nhau.

Sau khi thu được 8 bô chúng ta nối chúng lại với nhau và đem nhân với ma trận Wo lúc này chúng ta thu được model

Text

Description automatically generated with medium confidence

Trước khi cho dữ liệu để Self-Attention thì chúng ta có thể cho nó qua 1 phần dư đã

Diagram

Description automatically generated

Sau khi qua mã hóa chúng ta bắt đầu bước vào quá trình giải mã Bộ mã hóa bắt đầu bằng cách xử lý chuỗi đầu vào. Đầu ra của bộ mã hóa hàng đầu sau đó được chuyển đổi thành một tập hợp các vectơ chú ý K và V. Các vectơ này sẽ được sử dụng bởi mỗi bộ giải mã trong lớp “chú ý của bộ mã hóa-giải mã” giúp bộ giải mã tập trung vào những vị trí thích hợp trong chuỗi đầu vào:

Diagram

Description automatically generated

Dữ liệu sau khi qua bộ giải mã chúng ta cần đi qua hàm linear và softmax cuối cùng để có output:

Nhược điểm của mạng Transformer:

Độ phức tạp của mạng là O(T^2\*d) (với d là số embedding) trong khi trong các mô hình hồi quy độ phức tạp chỉ là O(T)

Với T các lớn trong các văn bản dài chi phí tính toán sẽ tăng lên rất lớn