TIỂU LUẬN HỌC PHẦN HỌC SÂU

ĐỀ TÀI:Ứng dụng Deeplearning trong tóm tắt văn bản Tiếng Việt

Sinh viên: Nguyễn Đình Việt Anh, Đỗ Mạnh Hùng, Đoàn Minh Hiển

K65A5 Khoa học dữ liệu
Khoa Toán - Cơ - Tin học
Trường Đại học Khoa học Tự nhiên - ĐHQGHN

Ngày 17 Tháng 12 Năm 2023

Mục lục

Mục lục

Trình bày về đề tài

Giới thiệu chung

Bài toán Text Summarization là bài toán tạo ra văn bản tóm tắt ngắn gọn, chính xác và trôi chảy cho một văn bản dài hơn.

Bài toán đặt ra

- Xây dựng mô hình tóm tắt cho văn bản Tiếng Việt.
- Xây dựng một giao diện cơ bản để người dùng truyền văn bản muốn tóm tắt

Hướng tiếp cận

- Tìm kiến dữ liệu văn bản liên quan và tiền xử lý
- Sử dụng Pretrained-model ViT5 và áp dụng phương pháp Fine-tuning
- Đánh giá trên văn bản thực

Mục lục

Mô tả về bộ dữ liệu

Nhóm sử dụng 3 bọ dữ liệu khác nhau

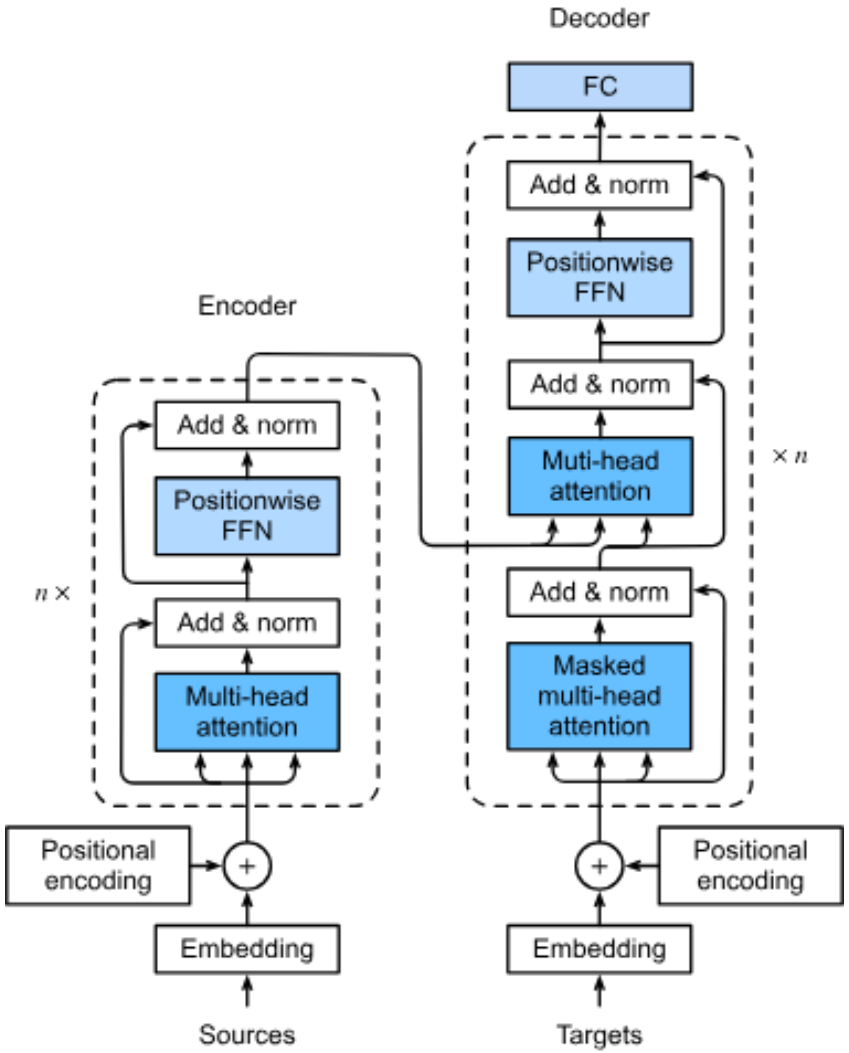
- Dataset ViMs: 300 cụm văn bản tiếng Việt của Nhóm tác giả ĐH KHTN Tp.HCM
- Datset UNK: 300k cụm văn bản bổ sung.

Mô tả dữ liệu

- Nguồn từ những trang báo nổi tiếng và phổ biến tại Việt Nam
- Ngoài ra còn 2 bộ dữ liệu khác được tổng hợp để củng cố thêm dữ liệu
- Đa dạng thể loại: Thế giới, Việt Nam, Kinh doanh, Giải trí, Thể thao
- Sau khi xử lý còn 2 cột: "Văn bản" và "Tóm tắt"
- Tổng số văn bản lên tới gần 300.000 mẫu văn bản

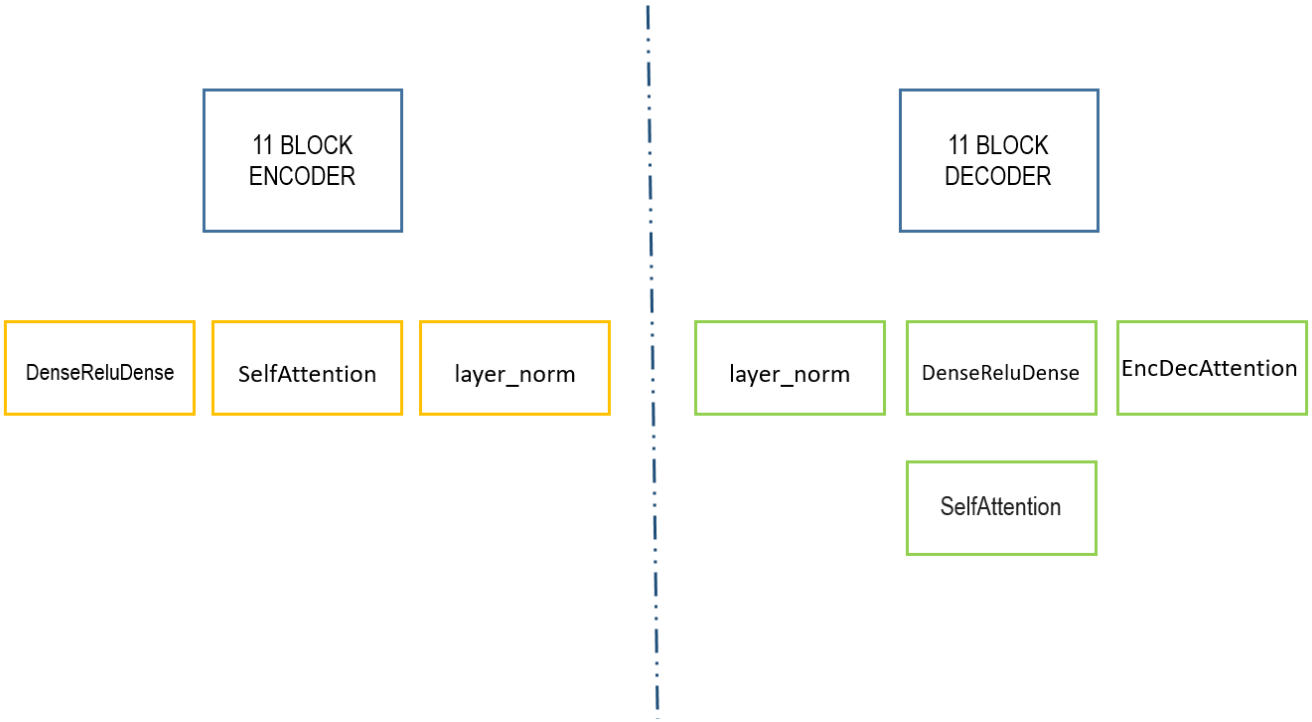
Mục lục

Cơ sở lý thuyết xây dựng mô hình



Hình 3.1: Kiến trúc Transformer

Cấu trúc mô hình



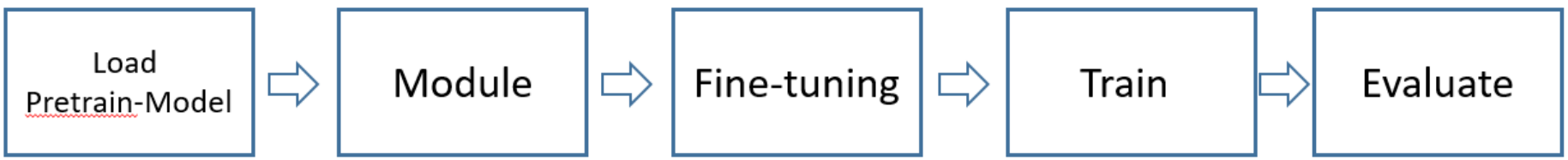
Hình 3.2: Enter Caption

Mục lục

Xây dựng mô hình

Các bước tiến hành

- Sử dụng AutoTokenize từ pretrained-model để xử lý dữ văn bản
- Xây dựng class Dataloader để quản lý dữ liệu
- Xây dựng classModelSummary để khởi tạo thông số đầu vào mô hình, và fine-tuning gồm tạo ModelCheckpoint, theo Dropout tại các layer, và tùy chỉnh learning rate
- Bắt đầu quá trình training và đánh giá



Hình 4.3: Work flow

Mục lục

Kết quả thực nghiệm

```
HIGH:
ROUGE1: Score(precision=0.6797322931404979, recall=0.5451093399288316, fmeasure=0.5778801369935239)
rouge2: Score(precision=0.3576956665465859, recall=0.2879295889782883, fmeasure=0.30421535649885795)
rougeL: Score(precision=0.46812873633416247, recall=0.3754622811545265, fmeasure=0.3976950532118971)
rougeLsum: Score(precision=0.46854875083458736, recall=0.37596938612410524, fmeasure=0.39830486499422024)

LOW:
ROUGE1: Score(precision=0.6605051168063049, recall=0.5252908929314305, fmeasure=0.5623952151507552)
rouge2: Score(precision=0.33411493030864586, recall=0.2668353638263977, fmeasure=0.28463738259559757)
rougeL: Score(precision=0.4485888281159517, recall=0.3560448802897342, fmeasure=0.38137739679658933)
rougeLsum: Score(precision=0.44794710830496826, recall=0.35676284190809604, fmeasure=0.38162531011927703)

MID:
ROUGE1: Score(precision=0.6701061600317633, recall=0.5352007071289606, fmeasure=0.5702874975539958)
rouge2: Score(precision=0.34522884306366586, recall=0.2773617764035904, fmeasure=0.29420258204857974)
rougeL: Score(precision=0.45822508932724604, recall=0.3658768559875526, fmeasure=0.38980495881849986)
rougeLsum: Score(precision=0.4584105355493352, recall=0.36589553700210986, fmeasure=0.3897438652765465)
```

Hình 5.4: ROUGE-Metric