NÂNG CAO TÓM TẮT TRÍCH XUẤT: CẢI TIẾN VÀ ĐỔI MỚI TRONG BERTSUM

Trần Hữu Lộc - 22520796
Trần Tuấn Khoa - 22520692

Tóm tắt

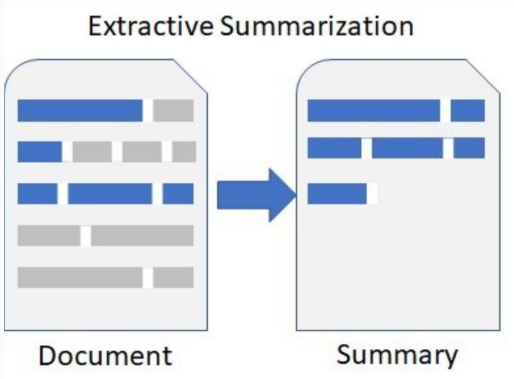
- Lớp: CS519.O21.KHTN
- Link Github: https://github.com/trhuuloc/CS519.O21.KHTN
- Link YouTube video: https://youtu.be/DH7wAiUT174
- Ảnh + Họ và Tên:

Trần Hữu Lộc - 22520796

Trần Tuấn Khoa - 22520692

Giới thiệu

- Tóm tắt trích xuất đóng vai trò quan trọng trong việc rút ngắn thông tin văn bản dài thành các bản tóm tắt ngắn gọn nhưng vẫn giữ lại được thông tin quan trọng. BERTSUM [1], sử dụng BERT [2], là 1 mô hình mạnh mẽ trong việc tạo ra các bản tóm tắt bằng cách tóm tắt trích xuất nhưng lại gặp khó khăn trong việc xử lý các phụ thuộc xa và quan hệ giữa các câu trong văn bản. Vì thế, chúng tôi thử nghiệm cải tiến và đổi mới BERTSUM.
- Ngoài ra, chúng tôi muốn tạo thêm bộ dữ liệu mới đủ tin cậy để phát triển vấn đề tóm tắt trích xuất văn bản.



Mục tiêu

- Tích hợp các mô hình transformer tiên tiến như RoBERTa [3] và Transformer-XL [4] trong BERTSUM cũng như sử dụng các kỹ thuật mã hóa dựa trên đồ thị để hiểu rõ hơn về mối quan hệ ngữ nghĩa giữa các câu. Đánh giá các kết quả và so sánh hiệu quả của BERTSUM sau khi áp dụng các cải tiến và đổi mới
- Tạo ra một chương trình minh họa áp dụng vào thực tiễn
- Tạo ra một bộ dữ liệu mới để cải thiện và phát triển các mô hình tóm tắt văn bản

Nội dung

Tìm hiểu câu trả lời cho câu hỏi: **“Liệu việc áp dụng tích hợp mô hình transformer tiên tiến như RoBERTa và Transformer-XL và sử dụng kỹ thuật mã hóa dựa trên đồ thị có nâng cao được hiệu quả cho bài toán tóm tắt trích xuất văn bản so với các phương pháp truyền thống và các mô hình tóm tắt trích xuất khác hay không?”**

Nội dung

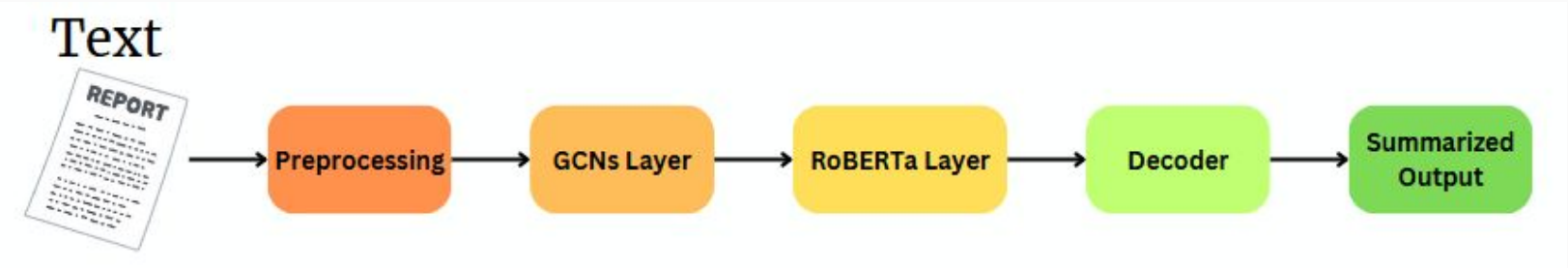
- **Tích hợp các mô hình transformer tiên tiến:** Nghiên cứu sử dụng và so sánh hiệu quả của các mô hình transformer như RoBERTa và Transformer-XL trong BERTSUM.
- **Sử dụng kỹ thuật mã hóa dựa trên đồ thị:** Nghiên cứu và phát triển các kỹ thuật để biểu diễn mối quan hệ ngữ nghĩa giữa các đoạn văn dưới dạng đồ thị.
- **Đánh giá và so sánh hiệu quả:** Nghiên cứu đánh giá các kết quả và so sánh hiệu quả của BERTSUM sau khi áp dụng các cải tiến so với các phương pháp truyền thống và các mô hình tóm tắt trích xuất khác trên độ đo ROUGE.
- **Tạo bộ dữ liệu mới đủ tin cậy để phát triển vấn đề tóm tắt trích xuất văn bản.**

Phương pháp

- **Thu thập và tiền xử lý dữ liệu:** Sử dụng các bộ dữ liệu phù hợp như CNN/DailyMail [6], XSUM [7] để huấn luyện và đánh giá BERTSUM, bao gồm việc tiền xử lý để chuẩn bị dữ liệu cho quá trình huấn luyện và kiểm thử. Ngoài ra, chúng tôi còn tự tạo ra 1 bộ dữ liệu mới đóng góp cho vấn đề tóm tắt trích xuất.
- **Huấn luyện và fine-tuning mô hình:**
  - Biểu diễn mối quan hệ ngữ nghĩa giữa các câu dưới dạng đồ thị.
  - Sử dụng GCNs [5] để mã hóa đồ thị ngữ nghĩa này, giúp nắm bắt các mối quan hệ phức tạp và phụ thuộc xa giữa các câu.
  - Huấn luyện các mô hình trên các bộ dữ liệu lớn như CNN/DailyMail và XSUM.
  - Fine-tuning các mô hình transformer như RoBERTa và Transformer-XL để tối ưu hóa hiệu suất của BERTSUM trong việc tóm tắt trích xuất.

Phương pháp

- **Đánh giá và so sánh kết quả:** Đánh giá khách quan hiệu quả của BERTSUM sau khi áp dụng các cải tiến và đổi mới, bao gồm so sánh với các mô hình tham chiếu và đo lường các tiêu chí như độ chính xác, độ bao phủ, và tính mạch lạc của bản tóm tắt.
- **Phân tích kết quả và đề xuất:** Phân tích các kết quả thu được từ nghiên cứu và đề xuất các hướng phát triển tiếp theo để cải thiện khả năng tóm tắt trích xuất của BERTSUM trong các ứng dụng thực tế.



Kết quả dự kiến

- BERTSUM có hiệu quả tốt hơn sau khi áp dụng các cải tiến và đổi mới với các phương pháp truyền thống và các mô hình tóm tắt trích xuất khác, từ đó đưa ra những đề xuất cụ thể để cải thiện hiệu suất và độ tin cậy của hệ thống tóm tắt.
- Áp dụng mô hình sau khi qua các cải tiến và đổi mới vào thực tiễn như tổng hợp tin tức, phân tích nội dung và hệ thống hỗ trợ quyết định.
- Có một bộ dữ liệu mới cho vấn đề tóm tắt trích xuất văn bản có độ tin cậy cao và được đưa vào sử dụng rộng rãi.

Tài liệu tham khảo

- [1]. Yang Liu: Fine-tune BERT for Extractive Summarization. CoRR abs/1903.10318 (2019)
- [2]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova:
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186
- [3]. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov: RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019)
- [4]. Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, Ruslan Salakhutdinov: Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. ACL (1) 2019: 2978-2988
- [5]. Thomas N. Kipf, Max Welling: Semi-Supervised Classification with Graph Convolutional Networks. ICLR (Poster) 2017
- [6]. Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, Bing Xiang: Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. CoNLL 2016: 280-290
- [7]. Shashi Narayan, Shay B. Cohen, Mirella Lapata:Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. EMNLP 2018: 1797-1807