# **THÔNG TIN CHUNG CỦA BÁO CÁO**

* Link YouTube video của báo cáo (tối đa 5 phút):

https://youtu.be/DH7wAiUT174

* Link slides (dạng .pdf đặt trên Github):

https://github.com/trhuuloc/CS519.O21.KHTN/blob/e398eef8b9b9227b35a8742cc8318c454e779eea/0796\_Tr%E1%BA%A7n%20H%E1%BB%AFu%20L%E1%BB%99c%20-%20CS519.O21.KHTN.DeCuong.FinalReport.Template.Slide.pdf

* *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
* *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

| * Họ và Tên: Trần Hữu Lộc * MSSV: 22520796 | * Lớp: CS519.O21.KHTN * Tự đánh giá (điểm tổng kết môn): 9/10 * Số buổi vắng: 0 * Số câu hỏi QT cá nhân: 5 * Link Github: <https://github.com/trhuuloc/CS519.O21.KHTN> * Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:   + Lên ý tưởng   + Viết nội dung   + Làm video Youtube   + Chỉnh sửa poster |
| --- | --- |
| * Họ và tên: Trần Tuấn Khoa * MSSV: 22520692 | * Lớp: CS519.O21.KHTN * Tự đánh giá (điểm tổng kết môn): 8,5/10 * Số buổi vắng: 1 * Số câu hỏi QT cá nhân: 5 * Link Github: [https://github.com/TKT314/CS519.O21.KHTN](https://github.com/../CS519.O21.KHTN) * Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:   + Làm slide   + Làm poster |

# **ĐỀ CƯƠNG NGHIÊN CỨU**

| **TÊN ĐỀ TÀI (IN HOA)**  NÂNG CAO TÓM TẮT TRÍCH XUẤT: CẢI TIẾN VÀ ĐỔI MỚI TRONG BERTSUM |
| --- |
| **TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**  ADVANCING EXTRACTIVE SUMMARIZATION: ENHANCING AND INNOVATIONS IN BERTSUM |
| **TÓM TẮT**  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 những nội dung quan trọng. Trong nghiên cứu này, chúng tôi xem xét các kỹ thuật tóm tắt trích xuất nâng cao, đặc biệt là các đổi mới trong BERTSUM [1].  BERTSUM, sử dụng mô hình BERT [2](Bi-Directional Encoder Representations from Transformers) mạnh mẽ, có khả năng tạo ra các bản tóm tắt bằng cách trích xuất các câu quan trọng trực tiếp từ văn bản nguồn.  Nghiên cứu này xem xét một số cải tiến để cải thiện hiệu quả tóm tắt của BERTSUM. Để xử lý hiệu quả hơn các phụ thuộc xa, các đổi mới đáng chú ý bao gồm việc tích hợp các mô hình transformer tiên tiến như RoBERTa [3] và Transformer-XL [4]. Ngoài ra, khả năng của mô hình xác định các mối quan hệ phức tạp giữa các câu được cải thiện khi tích hợp các kỹ thuật mã hóa dựa trên đồ thị. Điều này làm cho các bản tóm tắt được tạo ra trở nên mạch lạc hơn và chứa nhiều thông tin hơn.  Hơn nữa, chúng tôi đề xuất thêm một bộ dữ liệu mới với độ tin cậy cao cho vấn đề tóm tắt trích xuất văn bản nhằm mục đích tăng độ chính xác của mô hình, hỗ trợ nghiên cứu và phát triển sau này.  Tóm lại, những cải tiến được đề xuất cho BERTSUM là kết quả của sự tiến bộ đáng kể trong lĩnh vực trích xuất tóm tắt. Chúng hứa hẹn sẽ cải thiện đáng kể hiệu quả, độ chính xác và tính linh hoạt của các hệ thống tóm tắt tự động có nhiều ứng dụng. Nghiên cứu này đặt nền tảng cho những đổi mới trong tóm tắt trích xuất bằng cách tích hợp các kỹ thuật tiên tiến và tận dụng các kiến trúc transformer mạnh mẽ. Điều này sẽ nâng cao tối đa công nghệ tiên tiến trong lĩnh vực này. |
| **GIỚI THIỆU**  Các tóm tắt trích xuất giúp người dùng tiếp cận nhanh chóng các nội dung quan trọng mà không cần phải đọc toàn bộ tài liệu trong thế giới thông tin ngày nay. Về học tập và nghiên cứu, các bản tóm tắt cho phép sinh viên và nhà nghiên cứu nhanh chóng tìm ra các tài liệu phù hợp cho công việc của họ. Các ứng dụng đọc báo và tin tức sử dụng tóm tắt để cung cấp thông tin cho người dùng nhanh chóng và ngắn gọn, giúp cải thiện trải nghiệm người dùng.Tóm tắt trích xuất đóng vai trò quan trọng trong việc cung cấp phương pháp rút ngắn dữ liệu văn bản thành các bản tóm tắt ngắn gọn và thông tin. Khả năng này quan trọng trong các ứng dụng như truy xuất thông tin và phân tích nội dung, nơi mà việc tiếp cận hiệu quả đến thông tin liên quan là cực kỳ cần thiết.  BERTSUM là một bước tiến quan trọng trong các kỹ thuật tóm tắt trích xuất, được xây dựng trên nền tảng mạnh mẽ của mô hình BERT (Bi-Directional Encoder Representations from Transformers). BERTSUM tạo ra các bản tóm tắt rõ ràng và chứa đủ thông tin cần thiết bằng cách sử dụng các mô hình ngôn ngữ được huấn luyện trước.  Tuy nhiên, mặc dù hiệu quả, BERTSUM gặp khó khăn khi xử lý các phụ thuộc xa và hiểu các mối quan hệ phức tạp giữa các câu. Điều này thúc đẩy nghiên cứu tìm ra những cải tiến mới để nâng cao khả năng tóm tắt. Để giải quyết những vấn đề này và mở rộng khả năng của công nghệ tóm tắt trích xuất, nghiên cứu này tập trung vào các phương pháp tiên tiến trong khung BERTSUM.  Các cải tiến chính được khảo sát trong nghiên cứu này bao gồm tích hợp các kiến trúc transformer tiên tiến như RoBERTa và Transformer-XL. Những mô hình này được thiết kế để cải thiện khả năng xử lý và hiểu ngữ cảnh trên các chuỗi dài. Điều này sẽ giúp các bản tóm tắt được tạo ra trở nên mạch lạc hơn và chất lượng hơn.  Nghiên cứu cũng xem xét sự tích hợp các kỹ thuật mã hóa dựa trên đồ thị để hiểu rõ hơn về các mối quan hệ ngữ nghĩa giữa các câu. BERTSUM cho phép tạo ra các bản tóm tắt phản ánh rõ ràng hơn cấu trúc và ý nghĩa của nội dung gốc bằng cách trình bày các mối quan hệ văn bản dưới dạng đồ thị.  Nhận thấy các bộ dữ liệu về vấn đề tóm tắt trích xuất văn bản còn nhiều hạn chế và chưa đủ chất lượng và phong phú cũng như chưa đủ độ đa dạng, chúng tôi nhận thấy việc tạo ra một bộ dữ liệu mới là cần thiết nên chúng tôi đề xuất một bộ dữ liệu mới với mục đích mở rộng thêm dữ liệu và góp phần cải thiện độ chính xác cho vấn đề này.  Tóm lại, bài báo này trình bày một khám phá toàn diện về các cải tiến và đổi mới trong BERTSUM để tóm tắt trích xuất. Bằng cách tích hợp các kỹ thuật tiên tiến và tận dụng các mô hình transformer hiện đại, nghiên cứu này nhằm mở ra những tiềm năng cải tiến cho các hệ thống tóm tắt tự động, giúp cung cấp các bản tóm tắt chính xác và phản ánh ngữ cảnh một cách sâu sắc trong nhiều lĩnh vực ứng dụng khác nhau. |
| **MỤC TIÊU**   * Các mô hình như RoBERTa và Transformer-XL được áp dụng vào BERTSUM trong nghiên cứu để cải thiện khả năng xử lý các phụ thuộc xa và nắm bắt các mối quan hệ ngữ nghĩa phức tạp giữa các câu. Khảo sát và phát triển các phương pháp đồ thị hóa các mối quan hệ văn bản, từ đó nâng cao khả năng BERTSUM trong việc nắm bắt mối quan hệ ngữ nghĩa giữa các đoạn văn. Đánh giá kết quả của các cải tiến và đổi mới được áp dụng trong BERTSUM so với các phương pháp truyền thống và các mô hình tóm tắt trích xuất khác, nhằm đưa ra những đề xuất cụ thể để cải thiện hiệu suất và độ tin cậy của hệ thống tóm tắt. * Tạo ra một chương trình minh họa áp dụng vào thực tiễn với:   + Đầu vào: Văn bản dài cần được tóm tắt. Đây có thể là một bài viết, một đoạn văn, hoặc bất kỳ tài liệu văn bản nào.   + Đầu ra: Bản tóm tắt trích xuất. Bản tóm tắt cuối cùng được tạo ra bằng cách sắp xếp lại các câu được chọn theo thứ tự xuất hiện trong văn bản gốc. Các câu này được ghép lại để tạo thành một đoạn tóm tắt ngắn gọn và liên kết. * Tạo ra một bộ dữ liệu mới để cải thiện và phát triển các mô hình tóm tắt văn bản, giúp chúng trở nên chính xác hơn, phù hợp hơn với các lĩnh vực cụ thể và đáp ứng tốt hơn các nhu cầu của người dùng và thị trường. |
| **NỘI DUNG VÀ PHƯƠNG PHÁP**  **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?”**  Nghiên cứu này tập trung vào khám phá và đánh giá các cải tiến và đổi mới trong mô hình BERTSUM để nâng cao khả năng tóm tắt trích xuất. Ngoài ra, nghiên cứu này còn khai thác tạo ra một dữ liệu mới cho vấn đề tóm tắt trích xuất. Các nội dung chính của nghiên cứu này bao gồm:   1. **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. Điều này bao gồm việc huấn luyện và đánh giá các mô hình nhằm mục đích cải thiện khả năng xử lý các phụ thuộc xa và nắm bắt mối quan hệ ngữ nghĩa giữa các câu. 2. **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. Ở đây chúng tôi áp dụng GCNs [5] (Graph Convolutional Networks) để mã hóa đồ thị ngữ nghĩa này, kết hợp đầu ra của GCNs với BERTSUM để cải thiện khả năng nắm bắt mối quan hệ phức tạp giữa các câu. Điều này nhằm cải thiện khả năng của BERTSUM trong việc nắm bắt sự tương tác phức tạp giữa các câu và tạo ra các bản tóm tắt mạch lạc hơn. 3. **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 và đổi mới so với các phương pháp truyền thống và các mô hình tóm tắt trích xuất khác trên độ đo ROUGE. Điều này hỗ trợ trong việc xác định những cải tiến có hiệu quả cao nhất và có thể được áp dụng rộng rãi. 4. **Tạo bộ dữ liệu mới:** Tìm kiếm dữ liệu từ các bài báo, báo cáo, nghiên cứu hoặc bất kỳ tài liệu văn bản nào mà bạn cần tóm tắt. Bằng cách gắn nhãn các câu quan trọng trong mỗi tài liệu, bạn có thể tạo ra các bản tóm tắt. Sử dụng các phương pháp tự động, chẳng hạn như các mô hình tóm tắt hiện có, để tạo ra các nhãn ban đầu, hoặc quy trình này có thể được thực hiện thủ công bởi các chuyên gia.   **Phương pháp**   1. **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. 2. **Huấn luyện và fine-tuning mô hình:**    * Biểu diễn mối quan hệ ngữ nghĩa giữa các câu dưới dạng đồ thị.    * Sử dụng GCNs 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. 3. **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. 4. **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 MONG ĐỢI**   * BERTSUM sau khi áp dụng các cải tiến và đổi mới có hiệu quả tốt hơn so 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** *(Định dạng DBLP)*  [1]. [Yang Liu](https://dblp.org/pid/51/3710-124.html): Fine-tune BERT for Extractive Summarization. [CoRR abs/1903.10318](https://dblp.org/db/journals/corr/corr1903.html#abs-1903-10318) (2019)  [2]. [Jacob Devlin](https://dblp.org/pid/116/0575.html), [Ming-Wei Chang](https://dblp.org/pid/69/4618.html), [Kenton Lee](https://dblp.org/pid/121/7560.html), [Kristina Toutanova](https://dblp.org/pid/25/1520.html):  BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [NAACL-HLT (1) 2019](https://dblp.org/db/conf/naacl/naacl2019-1.html#DevlinCLT19): 4171-4186  [3]. [Yinhan Liu](https://dblp.org/pid/238/0128.html), [Myle Ott](https://dblp.org/pid/92/9767.html), [Naman Goyal](https://dblp.org/pid/183/1418.html), [Jingfei Du](https://dblp.org/pid/137/3917.html), [Mandar Joshi](https://dblp.org/pid/85/1261.html), [Danqi Chen](https://dblp.org/pid/87/7949.html), [Omer Levy](https://dblp.org/pid/117/4866.html), [Mike Lewis](https://dblp.org/pid/19/6214.html), [Luke Zettlemoyer](https://dblp.org/pid/21/6793.html), [Veselin Stoyanov](https://dblp.org/pid/275/3258.html): RoBERTa: A Robustly Optimized BERT Pretraining Approach. [CoRR abs/1907.11692](https://dblp.org/db/journals/corr/corr1907.html#abs-1907-11692) (2019)  [4]. [Zihang Dai](https://dblp.org/pid/182/1934.html), [Zhilin Yang](https://dblp.org/pid/54/6349.html), [Yiming Yang](https://dblp.org/pid/25/1666.html), [Jaime G. Carbonell](https://dblp.org/pid/56/3395.html), [Quoc Viet Le](https://dblp.org/pid/29/6166.html), [Ruslan Salakhutdinov](https://dblp.org/pid/62/5884.html): Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. [ACL (1) 2019](https://dblp.org/db/conf/acl/acl2019-1.html#DaiYYCLS19): 2978-2988  [5]. [Thomas N. Kipf](https://dblp.org/pid/186/8206.html), [Max Welling](https://dblp.org/pid/16/2286.html): Semi-Supervised Classification with Graph Convolutional Networks. [ICLR (Poster) 2017](https://dblp.org/db/conf/iclr/iclr2017.html#KipfW17)  [6]. [Ramesh Nallapati](https://dblp.org/pid/59/4797.html), [Bowen Zhou](https://dblp.org/pid/61/5024.html), [Cícero Nogueira dos Santos](https://dblp.org/pid/14/5278.html), [Çaglar Gülçehre](https://dblp.org/pid/125/2132.html), [Bing Xiang](https://dblp.org/pid/82/5456.html): Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. [CoNLL 2016](https://dblp.org/db/conf/conll/conll2016.html#NallapatiZSGX16): 280-290  [7]. [Shashi Narayan](https://dblp.org/pid/74/8458.html), [Shay B. Cohen](https://dblp.org/pid/04/5629.html), [Mirella Lapata](https://dblp.org/pid/59/6701.html):Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. [EMNLP 2018](https://dblp.org/db/conf/emnlp/emnlp2018.html#NarayanCL18): 1797-1807 |