


# THÔNG TIN CHUNG CỦA BÁO CÁO

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

*[https://github.com/zerokun218/ResearchMethodology/blob/main/DinhVanHoan\\_220101030\\_Slide\\_FinalReport.pdf](https://github.com/zerokun218/ResearchMethodology/blob/main/DinhVanHoan_220101030_Slide_FinalReport.pdf)*

<ul style="list-style-type: none"><li>● Họ và Tên: Đinh Văn Hoàn</li><li>● MSSV: 220101030</li></ul> 	<ul style="list-style-type: none"><li>● Lớp: CS2205.APR2023</li><li>● Tự đánh giá (điểm tổng kết môn): 7/10</li><li>● Số buổi vắng: 1</li><li>● Link Github: <i><a href="https://github.com/zerokun218/ResearchMethodology">https://github.com/zerokun218/ResearchMethodology</a></i></li><li>● Mô tả công việc:<ul style="list-style-type: none"><li>○ Lựa chọn đề tài</li><li>○ Viết báo cáo</li><li>○ Làm Slide</li><li>○ Làm Poster</li></ul></li></ul>
---	---

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

CẢI TIẾN TÀI LIỆU DẠNG VĂN BẢN BẰNG CÁCH ỨNG DỤNG MẠNG ĐỐI NGHỊCH TẠO SINH CÓ ĐIỀU KIỆN (CONDITIONAL GENERATIVE ADVERSARIAL NETWORK)

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

### TÓM TẮT *(Tối đa 400 từ)*

Với nhu cầu, mong muốn các bài luận, bài báo hay những đoạn văn mô tả ngắn được viết lại với cách diễn đạt hấp dẫn hơn, mạch lạc, sinh động và dễ dàng đọc hiểu, từ đó tiếp thu được những ý tưởng mà người viết muốn truyền đạt, tôi đề xuất sử dụng mạng đối nghịch tạo sinh có điều kiện (cGAN) để tăng cường tài liệu dạng văn bản. Trong đó, với các thông tin từ văn bản được đưa vào, mạng tạo sinh (Generator) sẽ chịu trách nhiệm cho việc tạo văn bản mới một cách ngẫu nhiên và mạng phân biệt (Discriminator) sẽ nhận biết xem văn bản được tạo ra ấy là thật hay giả. Chúng cạnh tranh lẫn nhau, huấn luyện nhau và từ đó cho ra mô hình có thể tạo ra được văn bản, tài liệu mới, chất lượng cao và thẩm mỹ, dễ tiếp cận và đủ để hấp dẫn, thu hút được sự tập trung của người đọc, do đó mở ra cánh cửa cho việc tạo ra tài liệu chất lượng.

Bằng cách ứng dụng mô hình trên, với nhu cầu mong muốn bài viết của mình trở nên hay hơn, người viết chỉ cần đưa vào văn bản của mình, dựa vào đó mà tạo ra một đoạn văn bản mới tuân thủ nội dung được đưa vào nhưng được biến tấu để làm cho câu văn trở nên sinh động, hấp dẫn hơn, cũng như vận dụng từ ngữ đa dạng để tránh gây sự nhàm chán khi đọc văn bản.

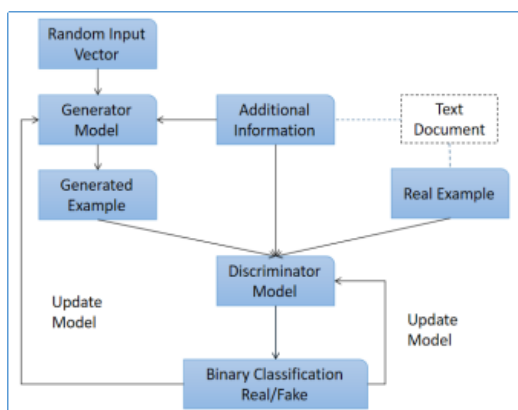
Tuy nhiên, với việc tạo ra văn bản mới, ta có thể sử dụng nhiều mô hình khác, đặc biệt là các mô hình thuộc dòng Transformer như BERT hay GPT. Các mô hình này có thể giải quyết khá tốt cho đề tài này, ví dụ như ChatGPT, nhưng điều mà tôi hướng đến là khả năng cạnh tranh lẫn nhau, làm đối thủ của nhau để hoàn thiện của

### **GIỚI THIỆU** *(Tối đa 1 trang A4)*

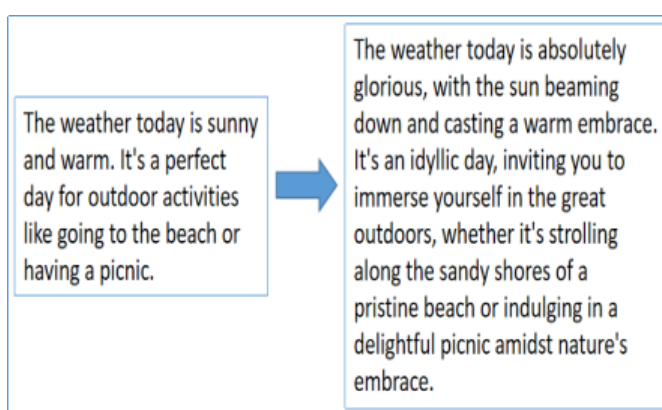
Trong thời đại hiện nay, khi mà lượng tài liệu con người có thể tiếp cận là rất lớn, chính vì thế mà con người thường ưu tiên việc tìm kiếm các tài liệu hữu dụng nhưng việc đọc hiểu các tài liệu đó cũng là một khó khăn và cần phải có kĩ năng để có thể khai thác được các thông tin hữu ích. Do đó mà nhu cầu nâng cao chất lượng và thẩm mỹ của tài liệu cũng trở nên quan trọng hơn bao giờ hết, đặc biệt là việc làm cho các văn bản trở nên hấp dẫn, sinh động, dễ tiếp cận và mang tính mạch lạc cao. Điều này giúp người đọc có thể dễ dàng tiếp cận nội dung mà người viết muốn bày tỏ, dễ dàng hơn trong việc khai thác nội dung. Nhưng với những người không có khả năng viết tốt để trình bày ý tưởng của mình thì lại là một khó khăn lớn.

Với những nhu cầu đó, tôi đề xuất ra phương pháp là ứng dụng mạng đối nghịch tạo sinh có điều kiện (Conditional Generative Adversarial Network - cGAN) để xây dựng nên mô hình tăng cường tài liệu văn bản.

Mạng cGAN gồm có 2 phần là mạng tạo sinh (Generator) và mạng phân biệt (Discriminator). Generator sẽ thực hiện việc tạo ra văn bản mới tuân thủ nội dung văn bản ban đầu và Discriminator sẽ thực hiện việc phân loại xem văn bản này có phải là thật hay là được tạo ra. Vì thế mà Generator và Discriminator được xem là đối thủ của nhau, Generator cố gắng tạo ra văn bản thật nhất để Discriminator không thể phân biệt được, cố gắng đánh lừa Discriminator và Discriminator cố gắng phân biệt tất cả những gì mà Generator cung cấp là thật hay giả. Bằng cách này, chúng thực hiện việc huấn luyện lẫn nhau, cho đến khi Generator có thể tạo ra văn bản thật nhất, giống với cách mà con người tạo ra văn bản.



Mô hình cGAN



Ví dụ

Đối với mạng GAN, Generator sẽ tạo ra một đoạn văn bản hoàn toàn ngẫu nhiên dựa vào một vector được khởi tạo một cách ngẫu nhiên. Điều này làm ta không thể kiểm soát được nội dung tạo ra. Chính vì thế mà một điều kiện (thông tin) sẽ được thêm vào để Generator hạn chế được phạm vi nội dung được tạo ra, văn bản tạo ra ấy phải đảm bảo nội dung tương tự với văn bản đưa vào. Sau khi Generator tạo được văn bản, Discriminator sẽ nhận vào văn bản này, dựa vào điều kiện (thông tin) được thêm vào và dữ liệu từ thực tế, từ đó tiến hành việc phân loại xem văn bản này là được tạo ra hay là văn bản thực tế. Dựa vào kết quả phân loại của Discriminator mà Generator cần phải hoàn thiện hơn để tạo ra mẫu mới trông thực hơn hay Discriminator cần phải học để phát hiện những đặc trưng mới cho việc phát hiện thật giả từ Generator. Chính quá trình cạnh tranh lẫn nhau của Generator và Discriminator mà ta có thể thu được một mô hình mà nó có thể tạo ra một văn bản hấp dẫn, linh động, mạch lạc như cách con người viết nhưng vẫn đảm bảo được nội dung ban đầu.

## MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

1. Xây dựng thành công mô hình cGAN để thực hiện việc tăng cường tài liệu dạng văn bản
2. Tính toán khả năng văn bản được tạo ra đảm bảo tuân thủ nội dung ban đầu
3. Ứng dụng được mô hình vào các trường hợp thực tế (Cải tiến bài báo, tiểu luận, nâng cao tính mạch lạc của bài văn)

## **NỘI DUNG VÀ PHƯƠNG PHÁP**

*(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)*

Để thực hiện việc huấn luyện mô hình cGAN cho văn bản, tôi sẽ sử dụng các tập dữ liệu dạng văn bản phổ biến là WebText và WikiText để thực hiện việc huấn luyện mô hình. Tuy nhiên, tôi vẫn cần phải có một tập dữ liệu chứa điều kiện (thông tin) để giới hạn nội dung mà Generator tạo ra. Chính vì thế, dựa trên nền tảng của 2 tập dữ liệu WebText, tôi dự định sẽ thực hiện việc rút gọn văn bản trên 2 tập này bằng cách loại bỏ những từ ngữ không quan trọng, không ảnh hưởng tới nội dung của câu nhưng điều này lại làm câu văn trở nên nghèo nàn, không còn mạch lạc và nhiệm vụ của Generator là làm cho các câu văn này hấp dẫn hơn, không còn khô khan nhưng vẫn đảm bảo tuân thủ nội dung gốc.

Sau khi mô hình đã được huấn luyện xong, tôi cần phải đánh giá xem liệu mô hình có tạo ra văn bản mà nội dung của nó không thay đổi hay không. Vì thế tôi sẽ tạo ra một tập test gồm 1000 đoạn văn bản cần được cải tiến để thực hiện kiểm tra. Sau khi Generator tạo ra 1000 văn bản mới tương ứng, tôi thực hiện việc đánh giá trên 1000 văn bản này liệu nó có nội dung tương tự với nội dung của văn bản ban đầu hay không.

## **KẾT QUẢ MONG ĐỢI**

*(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)*

Mô hình cGAN có thể cải tiến được văn bản mới trở nên hấp dẫn hơn, linh động và mạch lạc hơn, từ đó giúp người đọc dễ dàng tiếp cận cũng như tiếp thu được những ý tưởng mà người viết muốn trình bày.

Có khả năng ứng dụng mô hình vào các trường hợp thực tế như thực hiện cải tiến các bài tiểu luận, các đoạn văn ngắn, bài báo,...

## **TÀI LIỆU THAM KHẢO (Định dạng DBLP)**

[1]. de Rosa, Gustavo Henrique, and João Paulo Papa. "A Survey on Text Generation

Using Generative Adversarial Networks.” arXiv.Org, 20 Dec. 2022,  
[arxiv.org/abs/2212.11119](https://arxiv.org/abs/2212.11119).

[2]. Mirza, Mehdi, and Simon Osindero. “Conditional Generative Adversarial Nets.” arXiv.Org, 6 Nov. 2014, [arxiv.org/abs/1411.1784](https://arxiv.org/abs/1411.1784).

[3]. Goodfellow, Ian J., et al. “Generative Adversarial Networks.” arXiv.Org, 10 June 2014, [arxiv.org/abs/1406.2661](https://arxiv.org/abs/1406.2661).