

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

Đinh Văn Hoàn

Đại học Quốc gia Thành phố Hồ Chí Minh
Trường ĐH Công nghệ Thông tin

What ?

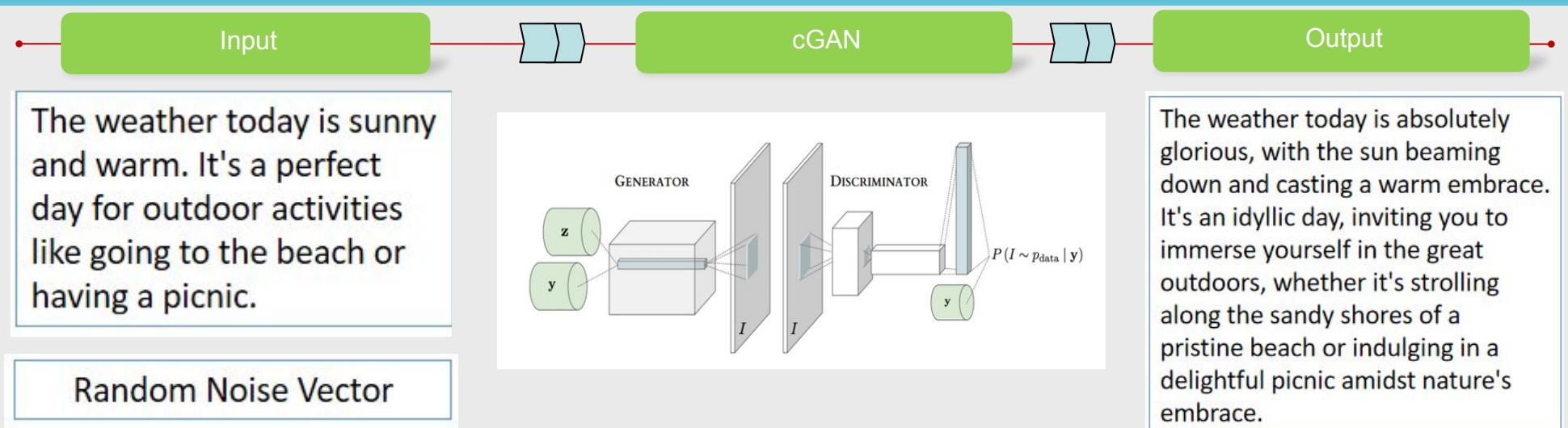
Sử dụng mô hình cGAN để thực hiện việc cải tiến tài liệu dạng văn bản

- Thực hiện cải tiến tài liệu văn bản nhưng vẫn đảm bảo nội dung ban đầu của văn bản, giúp văn bản trở nên hấp dẫn, thú vị, linh động và mạch lạc hơn
- Tận dụng khả năng cạnh tranh của Generator và Discriminator để nâng cao độ hiệu quả của mô hình

Why ?

- Tài liệu văn bản hiện nay có số lượng lớn, yêu cầu văn bản phải chất lượng để có thể đủ truyền đạt được ý tưởng mà tác giả muốn truyền đạt và cũng giúp người đọc dễ dàng tiếp cận. Tuy nhiên, không phải tất cả tác giả đều có khả năng viết tốt, đủ để có thể bày tỏ rõ ràng ý tưởng của mình.
- Mạng GAN hiếm khi được ứng dụng vào lĩnh vực tạo văn bản, chủ yếu ứng dụng vào việc tạo hình ảnh.

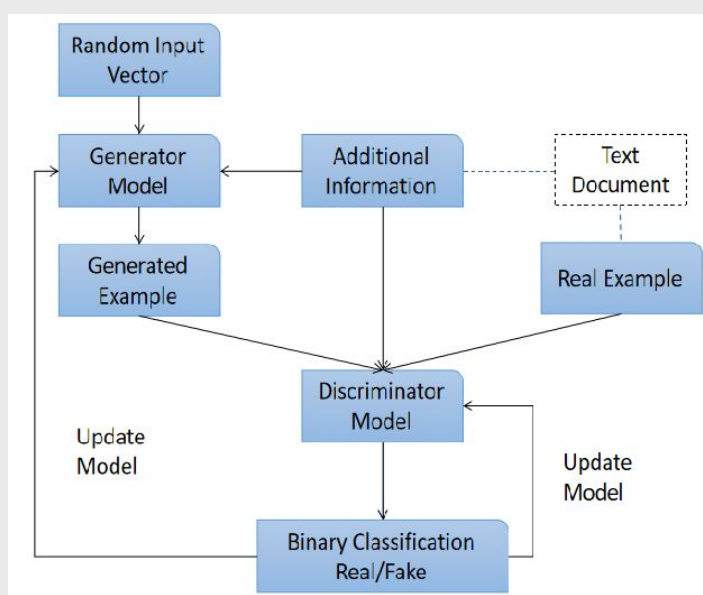
Overview



Description

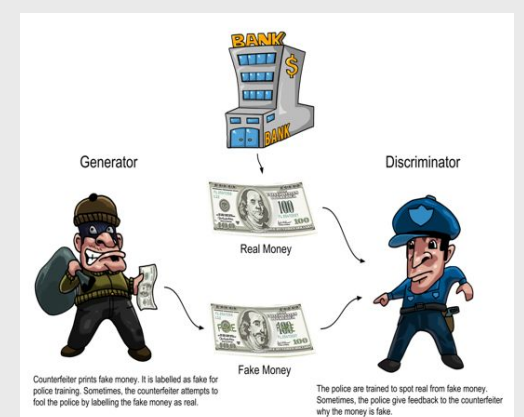
1. cGAN model

- Generator: chịu trách nhiệm cho việc tạo ra văn bản mới dựa trên điều kiện (thông tin) đưa vào để giới hạn phạm vi mà nó tạo ra
- Discriminator: chịu trách nhiệm cho việc phân biệt văn bản mà Generator cung cấp là thật hay giả



2. Generator vs Discriminator

- Generator và Discriminator đóng vai trò là đối thủ của nhau trong mạng GAN. Generator cố gắng tạo ra văn bản giống thực tế nhất có thể và Discriminator phải phân biệt được tất cả văn bản mà Generator cung cấp. Vì thế mà chúng tiến hành huấn luyện lẫn nhau, từ đó văn bản mà mô hình tạo ra được giống với cách mà con người viết.



2. Dataset

- Để huấn luyện mô hình, tôi huấn luyện mô hình trên 2 tập dữ liệu phổ biến:
 - WikiText: với hơn 100 triệu token được trích xuất từ các bài viết trên Wikipedia
 - WebText: an internal OpenAI corpus created by scraping web pages with emphasis on document quality

```
= Gold dollar =  
The gold dollar or gold one 8-8 dollar piece was a coin struck as a regular issue by the  
A gold dollar had been proposed several times in the 1830s and 1840s, but was not initi  
Gold did not again circulate in most of the nation until 1879; once it did, the gold d  
  
= Super Mario Land =  
Super Mario Land is a 1989 side 8-8 scrolling platform video game, the first in the Sup  
At Nintendo CEO Hiroshi Yamauchi 's request, Game Boy creator Gunpei Yokoi 's Nintendo  
Initial reviews were laudatory. Reviewers were satisfied with the smaller Super Mario B  
  
=== Sinclair Scientific Programmable ===  
The Sinclair Scientific Programmable was introduced in 1975, with the same case as the  
It had 24 8-8 step programming abilities, which meant it was highly limited for many pu  
However, included with the calculator was a library of over 120 programs that that perf
```