LÊ VĂN HIẾU – B19DCCN245

LỚP: D19CQCN05 - B

BỘ THÔNG TIN VÀ TRUYỀN THÔNG

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

ĐỒ ÁN

TỐT NGHIỆP ĐẠI HỌC

Đề tài: **Hierarchical Text-Conditional Image Generation System**

Giảng viên hướng dẫn:	**PGS. TS. Phạm Văn Cường**
Sinh viên thực hiện:	**Lê Văn Hiếu**
Lớp:	**D19HTTT2**
Khóa:	**2019 - 2024**
Hệ:	**Chính quy**

	Hà Nội – 2023

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN
TỐT NGHIỆP ĐẠI HỌC

Đề tài: **Hierarchical Text-Conditional Image Generation System**

Giảng viên hướng dẫn: **PGS. TS. Phạm Văn Cường**
Sinh viên thực hiện: **Lê Văn Hiếu**
Lớp: **D19HTTT2**
Khóa: **2019 – 2024**
Hệ: **Chính quy**

HÀ NỘI - 2023

LỜI CẢM ƠN

Lời đầu tiên, em xin bày tỏ lòng biết ơn chân thành tới thầy cô Khoa Công nghệ thông tin 1, Học viện Công nghệ Bưu chính Viễn thông, những người đã tận tâm giảng dạy và chia sẻ những kiến thức, kinh nghiệm quý giá trong suốt 4 năm qua, giúp em mở rộng tầm nhìn và chuẩn bị hành trang cho tương lai.

Em xin gửi lời cảm ơn sâu sắc đến PGS. TS. Phạm Văn Cường, người thầy hướng dẫn đã không quản ngại thời gian và công sức để hỗ trợ em trong suốt quá trình nghiên cứu và hoàn thành đồ án. Sự hướng dẫn tận tâm của Thầy đã giúp em tích lũy thêm nhiều kiến thức và kinh nghiệm quý giá.

Em xin trân trọng cảm ơn thầy Trần Tiến Công, người thầy cố vấn học tập trong suốt những năm trên giảng đường đại học. Thầy không chỉ là người truyền đạt kiến thức, mà còn là người hướng dẫn, chỉ bảo em trong từng bước đi, từ những bài học trong giáo trình cho đến những bài học quý báu trong cuộc sống. Sự kiên nhẫn, lòng nhiệt huyết và tâm huyết mà Thầy dành từng sinh viên, đã để lại trong em những ấn tượng sâu đậm và là nguồn động viên lớn trong mỗi bước em tiến trên con đường học vấn.

Em không quên gửi lời cảm ơn đến các anh chị, bạn bè, và các em tại các Câu lạc bộ, những người đã góp phần tạo nên một môi trường học tập năng động, sáng tạo, giúp em không ngừng phát triển. Cuối cùng, em xin gửi lời cảm ơn tới gia đình, bạn bè - những người luôn bên cạnh, hỗ trợ và tạo điều kiện tốt nhất cho em, giúp em có thể hoàn thành đồ án này.

Hà Nội, ngày 25 tháng 12 năm 2023

Tác giả

Lê Văn Hiếu

MỤC LỤC

DANH SÁCH BẢNG

DANH SÁCH HÌNH VẼ

DANH SÁCH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Tiếng Anh	Tiếng Việt/Giải thích
1	AI	Artificial Intelligence	Trí tuệ nhân tạo
2	ML	Machine Learning	Học máy
3	DNN	Deep Neural Network	Mạng nơron học sâu
4	CNN	Convolutional Neural network	Mạng nơron tích chập
5	LLM	Large language Model	Mô hình ngôn ngữ lớn
6	RNN	Recurrent Neural Network	Mạng nơron hồi quy
7	VAE	Variational Autoencoder	Mã hóa tự động biến thiên
8	GAN	Generative Adversarial Networks	Mạng đối nghịch tạo sinh
9	CLIP	Contrastive language-image pre-training	Đào tạo trước ngôn ngữ-hình ảnh tương phản

MỞ ĐẦU

Trong thế giới công nghệ số hiện đại, sự phát triển của trí tuệ nhân tạo (AI) đã mở ra những cánh cửa mới trong việc sáng tạo và tái hiện hình ảnh. Một trong những bước tiến đột phá nhất là khả năng sinh hình ảnh có điều kiện dựa trên mô tả văn bản.

Ngày có càng nhiều công cụ như Dalle, Midjourney, Stable Diffusion cho phép chúng ta biến những từ ngữ thành hình ảnh, mở ra không gian sáng tạo không giới hạn. Bằng cách sử dụng các mô hình học sâu, những công cụ này có thể phân tích và hiểu các mô tả văn bản, sau đó sinh ra hình ảnh phù hợp và độc đáo. Điều này không chỉ tạo ra những cơ hội mới cho các nhà thiết kế, nghệ sĩ, và nhà sáng tạo nội dung, mà còn giúp mọi người khám phá và thể hiện ý tưởng của mình một cách sinh động và chân thực hơn bao giờ hết. Ở phạm vi đồ án sẽ tập trung vào tìm hiểu các phương pháp và xây dựng mô hình có khả năng sinh hình ảnh có điều kiện dựa trên mô tả văn bản.

Tổng quan về nội dung của đồ án trình bày trong các chương sau, chi tiết như sau:

- **Chương 1 – Tổng quan:** Ở chương này, đồ án sẽ tập trung về các trí tuệ nhân tạo sáng tạo, giới thiệu phương pháp biểu diễn văn bản và hình ảnh, trình bày các nghiên cứu liên quan, từ đó giới thiệu các mô hình sinh và những kỹ thuật sử dụng để phục vụ cho bài toán sinh hình ảnh.

- **Chương 2 - Sinh hình ảnh có điều kiện theo mô tả văn bản:** Ở chương này, đồ án trình bày các cách để trích chọn các đặc trưng, biểu diễn các đặc trưng của hình ảnh và văn bản vào một không gian tiềm ẩn, sau đó kết hợp các mô hình sinh để tạo ra hình ảnh mới.

- **Chương 3 – Thực nghiệm và đánh giá:** Ở chương này, đồ án trình bày về lựa chọn bộ dữ liệu thích hợp, các phương pháp và chỉ số đánh giá của kết quả thu được trong những thực nghiệm đó.

Chương 1

Tổng quan

Trong chương 1, đồ án trình bày trí tuệ nhân tạo sáng tạo, những phương pháp biểu diễn văn bản và hình ảnh, trình bày những nghiên cứu và phương pháp sinh hình ảnh theo mô tả văn bản, giới thiệu về những mô hình sinh, từ đó phân tích và đề xuất phương pháp sinh hình ảnh có điều kiện theo mô tả văn bản qua các phần sau:

- Tổng quan về sinh hình ảnh theo mô tả văn bản
- Các nghiên cứu liên quan về sinh hình ảnh
- Mô hình sinh
- Mục tiêu đồ án

1.1 Tổng quan về sinh hình ảnh theo mô tả văn bản

1.1.1 Trí tuệ nhân tạo sáng tạo

Quá trình phát triển

Trí tuệ nhân tạo sáng tạo (Generative AI) là một loại hệ thống AI có khả năng tạo ra văn bản, hình ảnh hoặc các phương tiện truyền thông khác dựa trên các gợi ý (prompt). Những mô hình AI sáng tạo hầu hết áp dụng các kỹ thuật học máy mạng nơ-ron, sau đó tạo ra dữ liệu mới vẫn giữ được những đặc điểm then chốt cũng những đặc điểm sáng tạo được thêm vào. AI sáng tạo được giới thiệu vào những 1960s trong hệ thống chatbots. Ban đầu, các hệ thống này chủ yếu tập trung vào việc tạo ra văn bản dựa trên quy tắc cố định và mô hình ngôn ngữ đơn giản. Tuy nhiên, với sự tiến bộ của công nghệ và nghiên cứu, Generative AI ngày nay đã trở nên phức tạp và tinh vi hơn nhiều.

Với sự phát triển lĩnh vực học máy (machine learning) đã sử dụng các mô hình thống kế, mô hình sáng tạo để mô hình hóa và dự đoán, cuối những năm 2000, sự xuất hiện của học sâu (deep learning) đã thúc đẩy những tiến bộ và nghiên cứu trong xử lý ảnh và video, phân tích văn bản và các tác vụ khác. Nhưng phải đến năm 2014, những tiến bộ như autoencoder biến đổi (VAE) và mạng đội nghịch (GAN) đã tạo các mạng thần kinh thực tế đầu tiên có khả năng sáng tạo, thay vì phân biệt đối với các dữ liệu phức tạp như hình ảnh. Các mô hình đã có thể tạo ra các hình ảnh, video một cách chân thực thuyết phục được một bộ phận người sử dụng [1]. Hệ thống AI sáng tạo có thể là đơn mô-đun/phương thức (uni-modal) hoặc đa mô-đun/phương thức (multi-modal) có thể nhận nhiều loại đầu vào nhưng văn bản và hình ảnh là loại thường hay được sử dụng trong các hệ thống.

Văn bản

Văn bản là một loại hình phương tiện để ghi nhận, lưu giữ và truyền đạt các thông tin. Nó gồm tập hợp các câu có tính trọn vẹn về nội dung, hoàn chỉnh về hình thức, có tính liên kết chặt chẽ và hướng tới một mục tiêu giao tiếp nhất định. Hay nói khác đi, văn bản là một dạng sản phẩm của hoạt động giao tiếp bằng ngôn ngữ được thể hiện ở dạng viết trên một chất liệu nào đó. Trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), việc phân tích và hiểu văn bản đóng một vai trò trung tâm. Văn bản thường được phân cấp bao gồm một số lớp chi tiết, từ đơn vị nhỏ nhất đến đơn vị lớn nhất. Điều này cho phép máy tính không chỉ hiểu nghĩa đen của từng từ, mà còn hiểu cấu trúc và ý nghĩa tổng thể của cả văn bản.

Văn bản có thể được cấu trúc như sau:

- **Characters (Ký tự):** Cấp độ văn bản cơ bản nhất, thể hiện từng chữ cái, số, dấu câu và các ký hiệu kiểu chữ khác.
- **Tokens/Words (Token hoặc từ):** Cấp độ tiếp theo, trong đó các ký tự được nhóm thành mã, thường tương ứng với các từ hoặc đơn vị có ý nghĩa (như số hoặc chữ viết tắt).
- **Phrases (Cụm từ):** Các nhóm từ cùng nhau diễn đạt một phần ý tưởng hoặc khái niệm nhưng có thể không đứng riêng lẻ thành một câu hoàn chỉnh.
- **Sentences (Câu):** Một tập hợp các từ diễn đạt một ý nghĩ hoàn chỉnh và thường được giới hạn bởi cách viết hoa và dấu câu.
- **Paragraphs (Đoạn văn):** Các nhóm câu cùng nhau tạo thành một ý tưởng hoặc chủ đề.
- **Sections/Subsections (Mục hoặc tiểu mục):** Các phần lớn hơn của tài liệu chứa nhiều đoạn văn, thường được gắn nhãn bằng tiêu đề.
- **Chapters/Documents (Chương hoặc tài liệu):** Những phần hoàn chỉnh của tài liệu hoặc các phân chia chính của nó, chúng cũng có thể là một phần của các bộ sưu tập lớn hơn như series, tập sách, hoặc lưu trữ.
- **Corpora (Bộ sưu tập văn bản):** Các bộ sưu tập tài liệu liên quan được sử dụng cho việc phân tích trong các nhiệm vụ xử lý văn bản quy mô lớn.

Hình ảnh

Một bức hình, tấm ảnh, hay hình ảnh thứ ghi lại hay thể hiện/tái tạo được cảm nhận thị giác, tương tự với cảm nhận thị giác từ vật thể có thật, do đó mô tả được những vật thể đó. Hình ảnh có thể có hai chiều, như thể hiện trên tranh vẽ trên mặt phẳng, hoặc ba chiều, như thể hiện trên tác phẩm điêu khắc hoặc hologram. Hình ảnh có thể được ghi lại bằng thiết bị quang học như máy ảnh, gương, thấu kính, kính viễn vọng, kính hiển vi do con người tạo ra, hoặc bởi các cơ chế tự nhiên, như mắt người hay mặt nước.

Hình ảnh có thể được dùng theo nghĩa rộng, thể hiện bản đồ, đồ thị, nghệ thuật trừu tượng. Với nghĩa này, hình ảnh có thể được tạo ra mới hoàn toàn, thay vì ghi chép lại, bằng cách vẽ, tạc tượng, in ấn hay xây dựng bằng đồ họa máy tính. Hình ảnh tưởng tượng xuất hiện trong suy nghĩ của con người, tương tự như trí nhớ. Hình ảnh chuyển động có thể là phim, video, hoạt hình. Trong thời đại số hóa, công nghệ hình ảnh đã phát triển vượt bậc. Hình ảnh kỹ thuật số, được tạo ra và xử lý thông qua máy tính, đã trở thành một phần không thể thiếu trong cuộc sống hàng ngày. Chúng không chỉ được sử dụng trong lĩnh vực nghệ thuật, truyền thông, và giải trí, mà còn trong y học, khoa học, và công nghệ [2].

Nhìn chung những đặc điểm chính cấu tạo nên một hình ảnh như sau:

- **Màu sắc**: Một trong những đặc trưng quan trọng nhất của hình ảnh là màu sắc. Hệ thống cần có khả năng phân tích và hiểu mô tả về màu sắc từ văn bản, chẳng hạn như "đỏ rực", "xanh lá cây nhạt", hoặc "màu xám tro".

- **Hình dạng và kích thước**: Các đối tượng trong hình ảnh có thể có nhiều hình dạng và kích thước khác nhau. Hệ thống phải xác định và áp dụng những thông tin này, từ mô tả như "hình tròn lớn", "dài và mảnh", đến "nhỏ bé".

- **Bố cục và vị trí**: Vị trí của các đối tượng trong không gian hình ảnh cũng quan trọng. Mô tả văn bản có thể chỉ ra vị trí như "ở góc trên bên trái", "ở trung tâm", hoặc "phía sau cảnh trước".

- **Ngữ cảnh và môi trường**: Ngữ cảnh xung quanh và môi trường trong đó đối tượng xuất hiện cũng cần được xem xét. Điều này có thể bao gồm cảnh quan như "rừng nhiệt đới", "đường phố đông đúc", hoặc "phòng khách cổ điển".

- **Chất liệu và bề mặt**: Các đặc trưng về chất liệu và bề mặt như "bóng loáng", "nhám", hoặc "trong suốt" cũng cần được hiểu và tái hiện chính xác trong hình ảnh.

- **Ánh sáng và bóng tối**: Cách ánh sáng và bóng tối được mô tả trong văn bản cũng quan trọng, vì chúng ảnh hưởng đến cảm nhận tổng thể về hình ảnh. Các mô tả như "ánh sáng mờ ảo", "bóng tối u tối", hoặc "ánh nắng chói chang" cần được hiểu và tái tạo.

- **Biểu cảm và cảm xúc**: Trong trường hợp hình ảnh có sự hiện diện của con người hoặc sinh vật, biểu cảm và cảm xúc được mô tả cũng rất quan trọng, như "vui mừng", "buồn bã", hoặc "tò mò".

- **Chi tiết phụ và phong cách**: Cuối cùng, các chi tiết phụ và phong cách tổng thể của hình ảnh cũng cần được chú ý, như "phong cách cổ điển", "hiện đại", hoặc "surreal".

1.1.2 Các phương pháp biểu diễn hình ảnh

Biểu diễn hình ảnh là một khái niệm quan trọng trong xử lý ảnh và khoa học máy tính, liên quan đến cách hình ảnh được mã hóa và xử lý bởi máy tính. Có nhiều cách khác nhau để biểu diễn hình ảnh, mỗi cách phù hợp với mục đích cụ thể. Dưới đây là một số phương pháp biểu diễn hình ảnh phổ biến:

Biểu diễn dạng Pixel (Raster Graphics)

Biểu diễn dạng pixel, hay còn gọi là đồ họa raster, là một phương pháp phổ biến để biểu diễn hình ảnh trong máy tính. Đây là cách hình ảnh được mã hóa dưới dạng một lưới các pixel, mỗi pixel chứa thông tin về màu sắc và đôi khi độ sáng. Các định dạng hình ảnh raster thông dụng bao gồm JPEG, PNG, GIF, BMP, và TIFF.

Ưu điểm của biểu diễn:

- **Chi tiết và chất lượng hình ảnh cao**: Đồ họa raster có thể biểu diễn hình ảnh với chi tiết rất cao và độ chính xác màu sắc tốt, đặc biệt quan trọng cho hình ảnh phức tạp như ảnh chụp từ thiên nhiên, con người, hay nghệ thuật.

- **Chỉnh sửa dễ dàng**: Các hình ảnh raster có thể được chỉnh sửa dễ dàng với nhiều công cụ và phần mềm chỉnh sửa hình ảnh, cho phép sửa đổi từng pixel một nếu cần.

- **Tương thích rộng rãi**: Hình ảnh raster được hỗ trợ bởi hầu hết các ứng dụng và nền tảng web, làm cho chúng dễ dàng được xem và chia sẻ.

- **Phù hợp với hình ảnh phức tạp**: Do khả năng biểu diễn chi tiết cao, đồ họa raster phù hợp cho việc hiển thị các hình ảnh có độ phức tạp cao, như ảnh chụp, hình ảnh y khoa, và ảnh nghệ thuật

- **Hỗ trợ màu sắc đa dạng**: Hình ảnh raster có thể chứa hàng triệu màu sắc, giúp chúng rất phù hợp cho việc biểu diễn hình ảnh màu sắc phong phú và sống động.

Nhựu điểm của biểu diễn:

- **Kích thước file lớn**: Hình ảnh raster thường có kích thước file lớn do mỗi pixel chứa thông tin màu sắc riêng biệt. Kích thước file lớn có thể gây khó khăn trong việc lưu trữ và chia sẻ, đặc biệt qua mạng internet.

- **Mất chất lượng khi phóng to**: Khi hình ảnh raster được phóng to, các pixel trở nên rõ ràng hơn, dẫn đến hiện tượng "răng cưa" hoặc mờ. Điều này xảy ra bởi vì mỗi pixel chỉ biểu diễn một phần cụ thể của hình ảnh và không thể thích ứng với sự thay đổi kích thước.

- **Không thích hợp cho mọi ứng dụng**: Raster không phải lúc nào cũng là lựa chọn tốt nhất cho mọi ứng dụng. Trong các trường hợp cần đến độ chính xác cao hoặc khả năng thay đổi kích thước mà không làm mất chất lượng, đồ họa vectơ có thể là lựa chọn tốt hơn.

Biểu diễn Vectơ (Vector Graphics)

Trái ngược với raster, biểu diễn vectơ, hay đồ họa vectơ, là một phương pháp biểu diễn hình ảnh dựa trên các đường và hình dạng được xác định bởi các thuật toán toán học. Điều này tạo nên sự khác biệt cơ bản so với đồ họa raster, nơi mà hình ảnh được tạo nên từ các pixel. Các định dạng hình ảnh vectơ phổ biến bao gồm SVG, EPS, AI, và PDF.

Ưu điểm của biểu diễn:

- **Khả năng mở rộng không giới hạn**: Đồ họa vectơ có thể được phóng to hoặc thu nhỏ mà không làm mất chất lượng hình ảnh. Điều này là do các đường và

hình dạng được xác định bởi các phương trình toán học, không phụ thuộc vào số lượng pixel.

- **Kích thước file nhỏ**: Hình ảnh vectơ thường có kích thước file nhỏ hơn so với hình ảnh raster tương đương về độ phức tạp, làm cho chúng dễ dàng hơn trong việc lưu trữ và chia sẻ.

- **Dễ dàng chỉnh sửa**: Đồ họa vectơ dễ dàng được chỉnh sửa, với khả năng thay đổi màu sắc, kích thước, và hình dạng mà không ảnh hưởng đến chất lượng hình ảnh.

- **Lý tưởng cho đồ họa thiết kế**: Đồ họa vectơ rất phù hợp cho các thiết kế logo, biểu tượng, font chữ, và các yếu tố đồ họa khác mà cần độ chính xác cao và khả năng thích ứng với nhiều kích thước khác nhau.

- **Chất lượng in ấn cao**: Khi in ấn, đồ họa vectơ duy trì chất lượng cao, vì không bị ảnh hưởng bởi độ phân giải của máy in.

Nhựu điểm biểu diễn:

- **Khó khăn trong việc biểu diễn hình ảnh phức tạp**: Đồ họa vectơ không phù hợp cho việc biểu diễn hình ảnh chụp tự nhiên hoặc hình ảnh có độ phức tạp cao. Việc chuyển đổi hình ảnh tự nhiên sang định dạng vectơ có thể rất khó khăn và mất thời gian do sự phức tạp của các hình dạng và màu sắc.

- **Giới hạn tương thích với các ứng dụng**: Không phải tất cả các ứng dụng đều hỗ trợ hình ảnh vectơ, đặc biệt là trong lĩnh vực truyền thông kỹ thuật số và web. Một số ứng dụng và nền tảng có thể chỉ hỗ trợ hình ảnh raster, hạn chế khả năng sử dụng hình ảnh vectơ.

- **Thách thức trong việc hiển thị trên web**: Trong môi trường web, việc hiển thị và tương tác với đồ họa vectơ có thể gặp một số khó khăn do hạn chế về hỗ trợ và tối ưu hóa hiệu suất.

Biểu diễn màu sắc

Biểu diễn màu sắc trong xử lý hình ảnh và đồ họa máy tính là một yếu tố quan trọng, với nhiều cách biểu diễn khác nhau tùy thuộc vào mục đích sử dụng. Hình ảnh có thể được biểu diễn trong các không gian màu khác nhau như RGB (đỏ, xanh lá, xanh dương), HSV (màu sắc, độ bão hòa, giá trị độ sáng), và YCbCr (được sử dụng trong video và truyền hình).

RGB (Red, Green, Blue):

- **Ưu điểm**: Phản ánh cách mắt người nhận thức màu sắc, phù hợp cho hầu hết các ứng dụng hiển thị và xử lý hình ảnh.
- **Nhược điểm**: Không phản ánh độ sáng của màu sắc một cách trực tiếp, có thể khó xử lý trong một số ứng dụng như in ấn chuyên nghiệp.

CMYK (Cyan, Magenta, Yellow, Key/Black):

- **Ưu điểm**: Phù hợp cho in ấn, vì nó phản ánh cách mực in trộn lẫn với nhau.
- **Nhược điểm**: Không phù hợp cho màn hình hiển thị, hạn chế trong việc tái tạo một số màu sắc rực rỡ như màu RGB.

HSV (Hue, Saturation, Value):

- **Ưu điểm**: Biểu diễn màu sắc theo cách mà con người cảm nhận (màu sắc, độ bão hòa, và độ sáng), dễ dàng điều chỉnh các thành phần màu sắc.
- **Nhược điểm**: Không phổ biến như RGB trong xử lý hình ảnh kỹ thuật số và hiển thị.

LAB (Luminance, A, B):

- **Ưu điểm**: Phân biệt được màu sắc một cách chính xác và độc lập với thiết bị, tốt cho việc phân tích màu sắc.
- **Nhược điểm**: Không thông dụng trong các ứng dụng xử lý hình ảnh thông thường, phức tạp hơn trong việc hiểu và xử lý.

Mỗi không gian màu có những ưu và nhược điểm riêng, phù hợp với các ứng dụng cụ thể. Việc lựa chọn không gian màu phụ thuộc vào nhu cầu xử lý hình ảnh, hiển thị, và in ấn của dự án. RGB thường được sử dụng trong các ứng dụng hiển thị và xử lý

hình ảnh kỹ thuật số, trong khi CMYK phù hợp với in ấn, và HSV hay LAB thường được ứng dụng trong phân tích màu sắc chuyên nghiệp và nghiên cứu.

Biểu diễn dựa trên đặc trưng

Biểu diễn dựa trên đặc trưng (feature-based representation) trong xử lý hình ảnh và học máy là một phương pháp quan trọng, nơi hình ảnh được biểu diễn thông qua tập hợp các đặc trưng quan trọng thay vì sử dụng toàn bộ thông tin hình ảnh.

Ưu điểm của biểu diễn:

- **Giảm kích thước dữ liệu**: Biểu diễn dựa trên đặc trưng giúp giảm kích thước dữ liệu bằng cách loại bỏ thông tin không cần thiết, giữ lại chỉ những thông tin quan trọng.
- **Cải thiện hiệu suất xử lý**: Việc giảm kích thước dữ liệu giúp tăng tốc độ xử lý và giảm thời gian học của các mô hình học máy và xử lý hình ảnh.
- **Tăng khả năng phân loại và nhận dạng**: Các đặc trưng được chọn cẩn thận có thể cải thiện đáng kể khả năng phân loại và nhận dạng hình ảnh của các thuật toán.
- **Phù hợp với các ứng dụng cụ thể**: Biểu diễn dựa trên đặc trưng rất hiệu quả cho các ứng dụng cần nhận dạng mẫu hoặc phân tích hình ảnh cụ thể.

Nhưu điểm của biểu diễn:

- **Khó khăn trong việc chọn và tính toán đặc trưng**: Việc xác định và tính toán đặc trưng có thể phức tạp và đòi hỏi kiến thức chuyên môn.
- **Mất mát thông tin**: Khi loại bỏ các thông tin không quan trọng, có khả năng mất mát thông tin có thể quan trọng trong một số trường hợp.
- **Phụ thuộc vào lựa chọn đặc trưng**: Hiệu suất của phương pháp này rất phụ thuộc vào việc lựa chọn đúng các đặc trưng phù hợp và hiệu quả.

- **Không phổ quát**: Đặc trưng được chọn cho một loại hình ảnh hoặc tác vụ cụ thể có thể không phù hợp hoặc hiệu quả cho loại hình ảnh hoặc tác vụ khác.

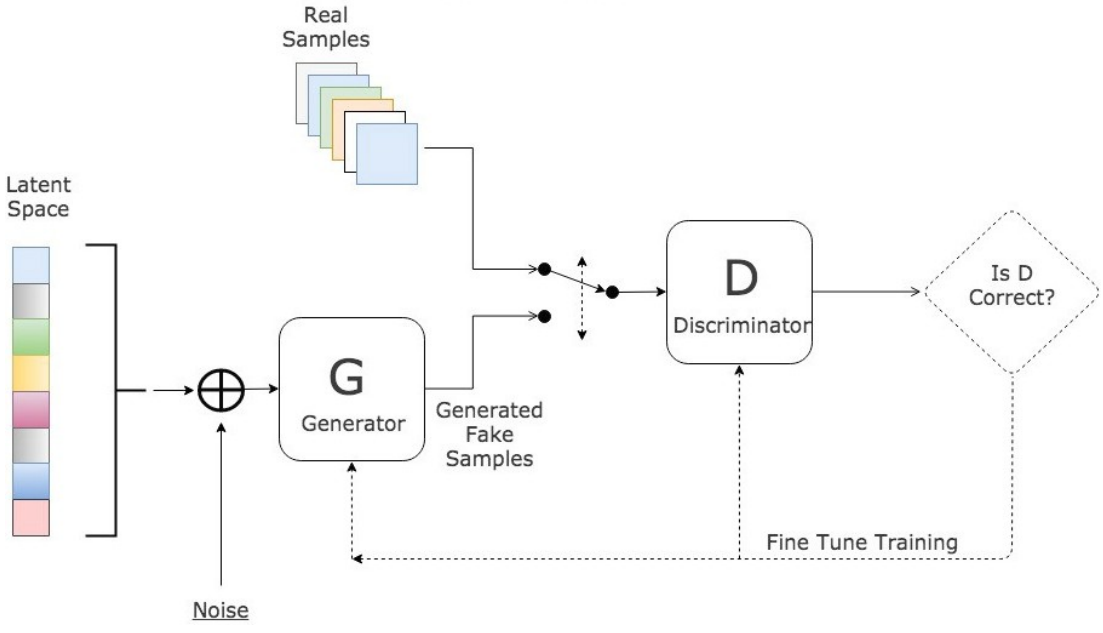
1.2. Các nghiên cứu liên quan về sinh hình ảnh

1.2.1. Tổng hợp văn bản thành hình ảnh sử Generative Adversarial Networks và bộ nhớ động (DM-GAN).

Nghiên cứu này trình bày hệ thống DM-GAN (Dynamic Memory Generative Adversarial Network) để tạo ra hình ảnh với chất lượng ảnh cao [3]. DM-GAN sử dụng một cấu trúc bộ nhớ động đặc biệt để cải thiện chất lượng và tính chính xác của hình ảnh được tạo ra từ mô tả văn bản. Trong đó GAN (Generative adversarial network) có vai trò tạo ra hình ảnh và Dynamic memory đảm bảo tính nhất quán của hình ảnh sinh và mô tả văn bản bằng cách ghi nhớ các đặc điểm quan trọng của hình ảnh cần được sinh ra, dựa trên mô tả văn bản.

Cấu trúc của GAN và cơ chế hoạt động

Generative adversarial network (GAN) là một mô hình học máy (machine learning) trong đó hai mạng thần kinh "đối đầu" với nhau bằng cách sử dụng các phương pháp học sâu (deep learning) để dự đoán của chúng ngày càng trở nên chính xác hơn. GAN sử dụng "khung trò chơi" để hai mạng cùng hợp tác tìm hiểu, trong đó lợi ích của một mạng bằng với tổn thất của mạng kia" kết quả là "trò chơi" được cải thiện có tổng bằng không [4].



Hình 1.1: Quá trình hoạt động trong GAN (Nguồn: https://www.kdnuggets.com).

Như Hình 1 .1 GAN được tạo nên từ hai mạng là mạng tạo (generator) và mạng phân biệt đối xử (discriminator).

- Generator là một mạng lưới thần kinh tích chập (convolutional neural network) la Mục tiêu của generator là tạo ra các kết quả đầu ra một cách nhân tạo có thể dễ bị nhầm lẫn với dữ liệu thực.
- Discriminator là một mạng lưới thần kinh giải mã (deconvolutional neural network). Mục tiêu của discriminator là xác định đầu ra nào mà nó nhận được được tạo ra một cách giả tạo.

Chi tiết về các bước để đào tạo GAN như sau:

- Generator sẽ xuất ra hình ảnh sau khi chấp nhận số ngẫu nhiên.
- Discriminator sẽ nhận được hình ảnh được tạo này cùng với luồng ảnh từ tập dữ liệu thực tế.
- Discriminator nhập cả hình ảnh thật và hình ảnh giả và đưa ra xác suất giá trị từ 0 đến 1 trong đó 1 biểu thị dự đoán về tính xác thực và 0 biểu thị giả mạo.
- Tạo ra một vòng phản hồi kép trong đó discriminator nằm trong vòng phản hồi với độ chân thực cơ bản của hình ảnh và generator nằm trong vòng phản hồi với bộ phân biệt.

Vấn đề khi gặp phải khi đào tạo là hội tụ GAN rất khó xác định. Khi generator cải thiện khả trong quá trình đào tạo, thì discriminator sẽ trở nên tệ đi vì vậy sẽ rất khó để nó có thể phân biện giữa giả mạo và thực tế. Nếu GAN tiếp tục đào tạo vượt qua thời điểm mà discriminator cung cấp phản hồi hoàn toàn ngẫu nhiên, thì generator sẽ bắt đầu đào tạo phản hồi "rác" và chất lượng của bộ phân biệt đó có thể bị thu gọn.

Cơ chế bộ nhớ tự động

Bộ nhớ động trong DM-GAN là một cấu trúc dữ liệu động, được thiết kế để lưu trữ, cập nhật và sử dụng thông tin liên quan trong quá trình tạo sinh hình ảnh. Nó hoạt động như một kho lưu trữ thông tin từ văn bản đầu vào và kết quả trung gian trong quá trình tạo sinh. Cách lưu trữ thông tin của bộ nhớ như sau:

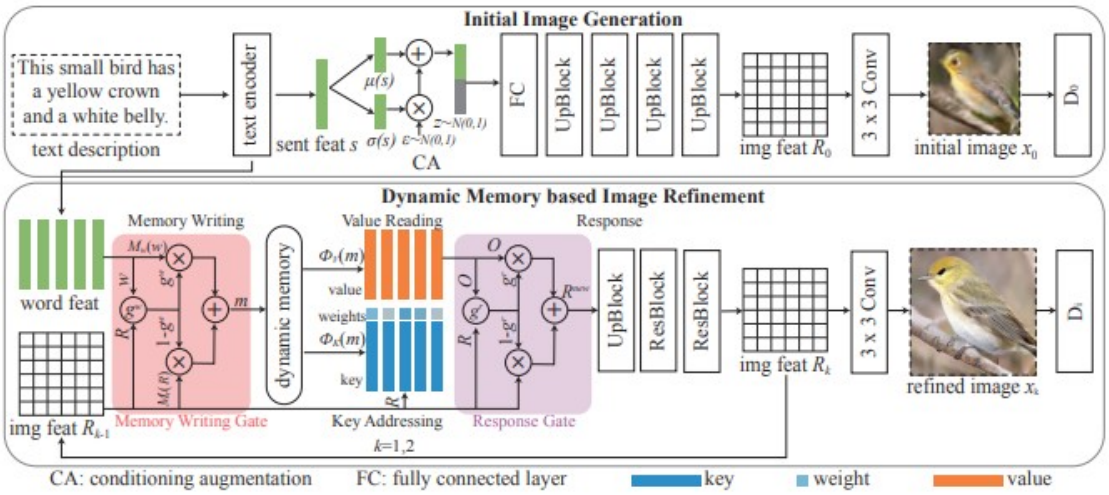
- Khi nhận văn bản đầu vào, bộ nhớ động tạo ra một biểu diễn vectơ của từng từ hoặc cụm từ chứa trong văn bản. Trong quá trình tạo sinh, bộ nhớ động cập

nhật liên tục dựa trên thông tin từ mạng tạo sinh. Mỗi khi mạng tạo sinh tạo ra một phần của hình ảnh, thông tin liên quan trong bộ nhớ động được cập nhật để phản ánh những thay đổi này.

- Các vectơ này được lưu trữ trong bộ nhớ, cho phép mô hình truy xuất nhanh chóng và hiệu quả khi cần thiết.

Nhờ cập nhật thông tin xuyên suốt quá trình, bộ nhớ động giúp DM-GAN kiểm soát chặt chẽ quá trình tạo, đảm bảo rằng hình ảnh tạo ra phản ánh chính xác nội dung của văn bản đầu vào. Thông tin được lưu trữ trong bộ nhớ động giúp mô hình tập trung vào các chi tiết quan trọng, từ đó tạo ra hình ảnh chính xác và chi tiết hơn.

Kiến trúc của DM-GAN



Hình 1.2: Kiến trúc DM-GAN để tổng hợp văn bản thành hình ảnh [3].

Theo như Hình 1 .2 DM-GAN gồm 2 giai đoạn: giai đoạn tạo văn bản ban đầu và giai đoạn sàng lọc hình ảnh dựa trên bộ nhớ động.

Giai đoạn tạo hình ảnh ban đầu [3], trước tiên, mô tả văn bản đầu vào được chuyển đổi thành một số biểu diễn bên trong (một đặc điểm câu s và một số đặc điểm từ W) bằng bộ mã hóa văn bản. Sau đó, một deep conventional generator dự đoán một hình ảnh ban đầu x0 có hình dạng thô và ít chi tiết theo đặc điểm câu và vectơ nhiễu ngẫu nhiên được tính bằng công thức sau:

$$z : x_0, R_0 = G_0(z, s) \qquad (1.1)$$

Trong đó R_0 là đặc điểm hình ảnh. Vectơ nhiễu được lấy mẫu từ phân phối chuẩn.

Giai đoạn sàng lọc hình ảnh dựa trên bộ nhớ động [3], nhiều nội dung hình ảnh chi tiết hơn được thêm vào các hình ảnh ban đầu mờ để tạo ra hình ảnh thực tế được biểu điễn bằng công thức sau:

$$x_i : x_i = G_i(R_{i-1}, W) \qquad (1.2)$$

Trong đó R_{i-1} là đặc điểm hình ảnh từ giai đoạn cuối cùng. Giai đoạn sàng lọc có thể được lặp đi lặp lại nhiều lần để lấy được thông tin thích hợp hơn và tạo ra hình ảnh có độ phân giải cao với nhiều chi tiết mịn hơn.

Kết quả thực nghiệm

Nghiên cứu được thực nghiệm trên 2 bộ dữ liệu CUB và COCO kết quả như trong Bảng 1 .1. Để đánh giá khả năng sinh hình ảnh thì bộ chỉ số thường được sử dụng đó là Inception Score (IS) và Frechet Inception Distance (FID). Các chỉ số đánh giá dựa trên 2 tiêu chí đó là:

- **Đa dạng nội dung của hình ảnh sinh ra** (Diversity): Một mô hình sinh tốt nếu tạo ra hình ảnh với sự đa dạng cao, không bị giới hạn trong một số ít khuôn mẫu. Điều này đảm bảo rằng mô hình có khả năng sinh ra nhiều loại hình ảnh khác nhau.

- **Tính xác thực của hình ảnh** (Realism): Hình ảnh được sinh ra phải trông thật, tức là chúng nên có đặc điểm giống với hình ảnh thực tế mà mô hình được huấn luyện trên. Điều này đánh giá khả năng của mô hình trong việc tạo ra hình ảnh có vẻ ngoài tự nhiên và hợp lý.

Trong đó mô hình có chỉ số IS càng cao tức là mô hình có khả sinh ảnh tốt (có một vài trường hợp thì IS không đánh giá đúng – chỉ đa dạng lớp chứ không đa dạng trong từng lớp) ngược lại FID là chỉ số không âm và càng thấp thì ảnh sinh ra càng giống như mô tả văn bản dựa trên dataset.

Dataset	**IS**	**FID**
CUB	4.75±0.07	16.09
COCO	30.49±0.57	32.64

Bảng 1.1: Kết quả thực nghiệm của hệ thống DM-GAN [3].

1.2.2. Tạo hình ảnh có điều kiện theo văn bản phân cấp với CLIP tiềm ẩn (DALL-E 2)

Nghiên cứu [5] DALL-E 2, được phát triển bởi OpenAI là một mô hình trí tuệ nhân tạo với đột phá về khả năng tạo ra hình ảnh phức tạp và cực kỳ chi tiết từ những mô tả văn bản. Nghiên cứu này tập trung vào việc cải thiện khả năng ngữ nghĩa của văn bản và biến đổi chúng thành hình ảnh. Bằng việc phối kết hợp với nhiều sử dụng nhiều kỹ thuật học sâu và mạng nơ ron như Tranformers [6], CLIP (Contrastive Language-Image Pre-Training) [7] và đặc biệt là mô hình Diffusion [8] hình ảnh được sinh ra từ mô hình mang rất nhiều ưu điểm so với các mô hình tại như khả năng hiểu biết ngữ nghĩa sâu sắc, tính linh hoạt và sáng tạo và khả năng tùy chỉnh.

DALL-E 2 tiến bộ trong lĩnh vực AI, minh họa cho khả năng sáng tạo và biến đổi của máy móc, đồng thời mở ra nhiều câu hỏi và thảo luận về tương lai của công nghệ này trong xã hội.

Kỹ thuật và phương pháp sử dụng

Quá trình trình huấn luyện của DALL-E 2 gồm hai giai đoạn đó là giai đoạn đầu huấn luyện CLIP (là mô hình học cách biểu diễn chung giữa văn bản và hình ảnh), giai đoạn sau CLIP sẽ được đóng băng để phục vụ cho quá trình sinh hình ảnh của Prior (có thể sử dụng AR và Diffusion) và Decoder (là mô hình đặc biệt dựa trên mô hình diffusion).

Ở giai đoạn CLIP sử dụng hai mạng mã hóa văn bản và hình ảnh thành không gian vector. Các mạng phục vụ cho qua trình này sẽ thường là ResNet (Residual neural network) hoặc ViT (Vision Transformer) cho hình ảnh và CBOW (Continuous bag of words) hoặc Text transformer cho văn bản. Sau đó CLIP được huấn luyện trên một tập dữ liệu lớn gồm cặp hình ảnh và văn bản. Mô hình học cách tối ưu hóa khoảng cách (trong không gian đặc trưng) giữa hình ảnh và văn bản tương ứng, đồng thời tăng khoảng cách với các cặp không tương ứng. Sau nhiều lần lặp lại quá trình tối ưu hóa CLIP từ đó sẽ có khả năng liên kết giữa văn bản và hình ảnh, tìm kiếm hình ảnh nâng cao, phân loại hình ảnh tự động, và hỗ trợ trong việc tạo ra hình ảnh từ mô tả văn bản. Nó cũng làm nổi bật tiềm năng của học máy và trí tuệ nhân tạo trong việc xử lý và hiểu dữ liệu phức tạp. Mô hình sau đó sẽ được đóng băng chuyển tiếp cho giai đoạn tiếp theo.

Giai đoạn tiếp theo, Prior xác định cách không gian đặc trưng ẩn được cấu trúc (không gian ẩn này chứa các biểu diễn của hình ảnh mà mô hình có thể sử dụng để sinh ra hình ảnh mới ở đây là CLIP), ngoài ra trong quá trình huấn luyện, Prior giúp hướng dẫn mô hình về cách nên biểu diễn dữ liệu. Điều này quan trọng để đảm bảo rằng mô hình học được cách sinh ra hình ảnh có ý nghĩa và chất lượng cao từ mô tả văn bản. Tóm lại thì Prior sẽ có vai trò kết nối ngữ nghĩa văn bản và ngữ nghĩa hình ảnh.

Sau khi quá trình Prior mô hình sẽ cho ra đâu ra là một hình ảnh Encoder và Decoder sẽ chuyển dổi chúng thành hình ảnh. Decoder phải tái tạo hình ảnh với chi tiết và độ phân giải cao, đảm bảo rằng hình ảnh sinh ra không chỉ chính xác mà còn có chất lượng hình ảnh tốt. Ngoài ra, Decoder trong DALL-E 2 cũng phải linh hoạt và sáng tạo để có thể sinh ra hình ảnh trong nhiều phong cách khác nhau và phản ứng với nhiều loại mô tả văn bản.

Kết quả thực nghiệm

Ở nghiên cứu này mô hình thử nghiệm so sánh khi sử dụng 2 mô hình trong Prior là Autoregresive (AR) và Diffusion. Nghiên cứu sẽ được tiến hành đánh giá bằng con người (Human Evaluations) một cách có hệ thống so sánh về tính chân thực (Photorealism), độ tương tự của chú thích (Caption Similarity) và tính đa dạng của mẫu (Diversity) [5].

unCLIP Prior	Photorealism	Caption Similarity	Diversity
AR	47.1% ± 3.1%	41.1% ± 3.0%	62.6% ± 3.0%
Diffusion	48.9% ± 3.1%	45.3% ± 3.0%	70.5% ± 2.8%

Bảng 1.2: Đánh giá của con người unCLIP prior [5].

Như Bảng 1 .2, nhìn chung Diffusion hoạt động tốt hơn so với AR. Ngoài ra, nghiên cứu này thực hiện các cuộc khảo sát thì kết quả mọi người vẫn thích GLIDE (Guided Language to Image Diffusion for Generation and Editing) [9] hơn unCLIP về mặt quang học, nhưng không đáng kể. Thậm chí với tính chất quang học tương tự, unCLIP vẫn được ưa chuộng hơn GLIDE xét về tính đa dạng và nổi bật một trong những lợi ích của nó.

Ngoài ra, nghiên cứu còn so sánh với các mô hình khác dựa trên chỉ số FID sử dụng bộ dữ liệu MS-COCO (256 × 256) thì cho thấy mô hình được đánh giá khá tốt [5].

1.3. Mô hình sinh

Mô hình sinh (Generative model) trong học máy (Machine learning) và thống kê là loại mô hình được xây dựng để tự động học cách tạo ra dữ liệu mới có đặc điểm tương tự như tập dữ liệu mà nó được huấn luyện. Các mô hình này không chỉ dự đoán kết quả (như mô hình phân loại hoặc hồi quy) mà còn có khả năng sinh ra các mẫu dữ liệu mới, có thể dùng để hiểu sâu hơn về cấu trúc và quy luật tiềm ẩn trong dữ liệu. Mở ra, những hướng đi mới trong lĩnh vực Vision Computer trong tương lai.

Các nhà nghiên cứu đã phát hiện ra tiềm năng của các mô hình AI sáng tạo mới vào những năm 2014, 2015 khi các autoencoder biến đổi (VAE), mạng đối nghịch tạo sinh (GAN) và các mô hình khuếch tán (Diffusion) được phát triển. Máy biến áp (Transformers), mạng nơ-ron đã có những đột phá có thể phân tích trên các bộ dữ liệu lớn với quy mô lớn để tự động tạo các mô hình ngôn ngữ lớn (LLM), được ra mắt vào năm 2017 [1].

Mỗi mô hình đều có những điểm mạnh và điểm yếu riêng trong từng trường hợp bài toán đặt ra. Xét đến bây giờ, các mô hình Diffusion đang được đánh giá hoạt động rất tốt trong lĩnh vực tổng hợp hình ảnh và video điển hình như Stable diffusion, Dalle-2 …, Transformer thì lại vượt trội trong lĩnh vực văn bản. GAN phù hợp trong việc tăng cường các tập dữ liệu nhỏ bằng các mẫu tổng hợp. Việc lựa chọn những mô hình nào tốt nhất luôn tùy thuộc vào trường hợp sử dụng cụ thể, yêu cầ chúng ta phải nắm ra rõ được cách các mô hình hoạt động ra sao ưu nhữu điểm của từng mô hình. Mỗi một mô hình sinh sử dụng các kỹ thuật và cách tiếp cận khác nhau để tạo ra dữ liệu mới như sau:

Variational Autoencoders

Variational Autoencoders (VAEs), được phát triển vào năm 2014, là một phương pháp tiên tiến sử dụng mạng nơ-ron để mã hóa dữ liệu một cách hiệu quả. VAEs được thiết kế để cải thiện cách biểu diễn thông tin. Cấu trúc của chúng bao gồm hai thành phần chính: một bộ mã hóa giúp giảm kích thước dữ liệu và một bộ giải mã tái tạo dữ liệu về trạng thái ban đầu của nó. Các VAEs rất phù hợp cho việc tạo ra dữ liệu mới từ dữ liệu đã được mã hóa, cải thiện hình ảnh hoặc dữ liệu bị nhiễu, phát hiện bất thường, và điền vào các thông tin còn thiếu [10].

Tuy nhiên, VAEs cũng có một số hạn chế, chẳng hạn như xu hướng tạo ra hình ảnh không rõ nét hoặc chất lượng kém. Thách thức khác liên quan đến không gian ẩn – một không gian có số chiều thấp được sử dụng để mô phỏng cấu trúc của dữ liệu, thường khá phức tạp và khó để xử lý. Những hạn chế này có thể làm giảm hiệu quả của VAEs trong các ứng dụng đòi hỏi hình ảnh chất lượng cao hoặc sự hiểu biết chi tiết về không gian ẩn. Do đó, các phiên bản tiếp theo của VAEs có thể sẽ tập trung vào việc cải thiện chất lượng dữ liệu sinh ra, tăng tốc quá trình huấn luyện và khai thác tiềm năng ứng dụng trong việc xử lý dữ liệu tuần tự [10].

Generative Adversarial Networks

Generative Adversarial Networks (GANs), ra đời vào năm 2014, là một kỹ thuật tiên tiến trong lĩnh vực học máy, ban đầu được thiết kế để sinh ra khuôn mặt người và số liệu một cách thuyết phục. GAN hoạt động dựa trên nguyên lý của hai mạng nơ-ron: một mạng "sinh" (generator) tạo ra dữ liệu, và một mạng "phân biệt" (discriminator) có nhiệm vụ phân biệt dữ liệu thật và giả. Qua quá trình huấn luyện cạnh tranh giữa hai mạng này, GANs học cách tạo ra hình ảnh không thể phân biệt được với dữ liệu thực [11].

GANs được ứng dụng rộng rãi trong việc tạo hình ảnh, chỉnh sửa ảnh, tăng cường độ phân giải, tạo dữ liệu giả mạo cho huấn luyện, chuyển đổi phong cách nghệ thuật, sản xuất âm nhạc, và tạo nội dung deepfake.

Một trong những thách thức chính của GANs là vấn đề về sự sụp đổ chế độ (mode collapse), nơi mạng sinh chỉ tạo ra một số lượng hạn chế các mẫu, làm giảm đa dạng của dữ liệu sinh ra. Điều này cũng gây khó khăn trong quá trình huấn luyện. Thế hệ tiếp theo của GANs đang tập trung vào việc cải thiện sự ổn định và hội tụ trong quá trình đào tạo, mở rộng ứng dụng sang các lĩnh vực mới và phát triển các phương pháp đánh giá hiệu quả hơn. Ngoài ra, GANs cũng được biết đến với tính khó tối ưu hóa và thiếu sự kiểm soát rõ ràng đối với mẫu được sinh ra, đặt ra những thách thức trong việc ứng dụng chúng một cách hiệu quả.

Diffusion

Mô hình khuếch tán (Diffusion Model), ban đầu được phát triển bởi một nhóm nghiên cứu tại Đại học Stanford vào năm 2015, là một phương pháp mô hình hóa động lực học entropy và nhiễu trong dữ liệu. Phương pháp này, nổi bật nhờ sự phát triển của

ứng dụng Stable Diffusion vào năm 2022, đã thu hút sự chú ý đến kỹ thuật khuếch tán, một kỹ thuật đã có từ lâu. Các mô hình khuếch tán này giúp mô phỏng quá trình như cách muối khuếch tán trong nước và sau đó đảo ngược quá trình đó, và chúng cũng rất hiệu quả trong việc tạo ra nội dung mới từ một hình ảnh ban đầu trống không [8].

Hiện tại, mô hình khuếch tán đang dẫn đầu trong lĩnh vực tạo hình ảnh. Chúng là nền tảng cho nhiều dịch vụ sinh hình ảnh nổi tiếng như Dall-E 2, Stable Diffusion, Midjourney và Imagen. Không chỉ giới hạn ở hình ảnh, những mô hình này còn được áp dụng trong việc tạo ra giọng nói, video và nội dung 3D. Hơn nữa, phương pháp khuếch tán cũng thích hợp để xử lý và dự đoán dữ liệu thiếu. Trong nhiều ứng dụng hiện đại, mô hình khuếch tán được kết hợp với các mô hình ngôn ngữ lớn (LLM) để chuyển đổi văn bản thành hình ảnh hoặc video. Một ví dụ điển hình là Stable Diffusion 2, sử dụng mô hình đào tạo trước ngôn ngữ-hình ảnh tương phản (CLIP) làm bộ mã hóa văn bản và bổ sung thêm các mô hình khác để tăng cường chiều sâu và chất lượng.

Transformers

Transformers, ra mắt vào năm 2017 bởi Google Brain, được thiết kế ban đầu để tăng cường khả năng dịch ngôn ngữ. Công nghệ này đặc biệt hiệu quả trong việc xử lý dữ liệu một cách không tuần tự, cho phép xử lý song song và mở rộng quy mô lên các mô hình lớn mà không cần dữ liệu có nhãn [6]. Công nghệ Transformer có nhiều ứng dụng rộng rãi, từ tóm tắt văn bản, xây dựng chatbots, hệ thống đề xuất, dịch thuật, cơ sở dữ liệu tri thức, đến việc cá nhân hóa sâu qua mô hình dựa trên sở thích. Nó cũng được sử dụng trong phân tích cảm xúc, nhận dạng thực thể như con người, địa điểm và sự kiện, nhận dạng giọng nói (như các mô hình của OpenAI), phát hiện đối tượng trong video và hình ảnh, gắn nhãn hình ảnh, phân loại văn bản và tạo hội thoại.

Dù có khả năng linh hoạt và đa dạng, Mô hình Transformers, đặc biệt là các phiên bản quy mô lớn như GPT-3, yêu cầu một lượng lớn tài nguyên tính toán để đào tạo. Điều này không chỉ đòi hỏi phần cứng mạnh mẽ như GPU và TPU chuyên dụng, mà còn gây ra chi phí vận hành cao. Để đạt được hiệu suất tốt nhất, các mô hình Transformers cần được đào tạo trên bộ dữ liệu lớn. Các thách thức này cần được giải quyết để tối ưu hóa hiệu quả và độ phổ biến của công nghệ này.

Flow-based Model

Khái niệm về mô hình Flow-based bắt nguồn từ ý tưởng về các biến đổi đảo ngược và việc mô hình hóa phân phối dữ liệu. Các nghiên cứu đầu tiên liên quan đến việc áp

dụng các biến đổi toán học để chuyển đổi dữ liệu từ một phân phối phức tạp sang một phân phối đơn giản hơn, và ngược lại, xuất hiện từ những năm cuối của thập kỷ 1990 và đầu thập kỷ 2000. Các nghiên cứu này tập trung vào việc áp dụng các biến đổi toán học để chuyển đổi dữ liệu từ phân phối phức tạp sang phân phối đơn giản hơn, mở đường cho việc phát triển các mô hình Flow-based hiện đại. Điều này đồng nghĩa với việc có thể dễ dàng chuyển từ dữ liệu sinh ra trở lại với dữ liệu ban đầu. Bằng cách thực hiện thông qua tính toán ma trận Jacobian của các biến đổi [12].

Mô hình Flow-based cung cấp sự linh hoạt cao và khả năng sinh dữ liệu chất lượng tốt. Tuy nhiên, chúng có thể yêu cầu nhiều tài nguyên tính toán và bộ nhớ, đặc biệt khi xử lý với dữ liệu có kích thước lớn.

1.4. Mục tiêu của đồ án

Mục tiêu đồ án là xây dựng hệ thống sinh hình ảnh theo văn bản phân cấp với hình ảnh vẫn giữ được nội dung chính của văn bản cung cấp. Ngoài ra, hình ảnh được tạo ra phải thực sự đa dạng, không chỉ đảm bảo về mặt nội dung mà còn thuyết phục về mặt thẩm mỹ, hệ thống đó là sử dụng mô hình sinh kết hợp với các mạng nơ-ron, không gian ẩn trên tập dữ liệu chất lượng và lớn. Ưu điểm của hệ thống có khả năng đáp ứng những yêu cầu như đã đề ra, khả năng mở rộng cao và phần nào kiểm soát được hình ảnh sinh ra theo hướng của mình. Tuy nhiên, đây là một hệ thống thực sự phức tạp với lượng tham số có thể lên đến hàng tỷ, việc xử lý trên tập dữ liệu khá lớn để có thể huấn luyện được mô hình là một thách thức lớn về cả trang thiết bị lẫn thời gian. Những năm gần đây với sự phát triển của các mô hình sinh, cộng đồng phát triển mạnh, vì vậy có tái sử dụng các mô hình cần trong hệ thống đã tin cậy phần nào giúp giảm bớt chi phí trong thời gian huấn luyện. Đồ án kỳ vọng sẽ thành công xây dựng được mô hình đáp ứng những mong muốn đã đề ra.

Chương 2

Sinh hình ảnh có điều kiện theo mô tả văn bản

Ở chương 2, đồ án sẽ trình bày chi tiết về cách phối hợp giữa mô hình sinh và không gian tiềm ẩn cho hệ thống sinh hình ảnh có điều kiện theo mô tả văn bản quá các phần sau:

- Tiền xử lý dữ liệu
- Trích chọn đặc trưng
- Biểu diễn hình ảnh và văn bản trong không gian tiềm tiềm ẩn
- Huấn luyện mô hình sinh cho phép tạo hình ảnh từ văn bản

2.1. Tiền xử lý dữ liệu

Đối với những bài toán sinh hình ảnh từ mô tả văn bản, thì dữ liệu cần để phục vụ cho đồ án là một cặp gồm hình ảnh và mô tả văn bản cho hình ảnh đó. Dữ liệu này sẽ thường được lưu dưới định dạng như sau:

Dữ liệu văn bản

Dữ liệu văn bản trên cộng đồng có thể từ sách, bài báo, bài đăng trên mạng xã hội, và nhiều nguồn khác nữa. Dữ liệu này thường là dữ liệu thô, trong chuỗi văn bản còn chứa khá nhiều đoạn nội dung gây nhiễu không mang thông tin gì cũng như không mô tả ý nghĩa của hình ảnh đi kèm. Việc xử lý dữ liệu văn bản là quá trình khá tốn thời gian, quá trình này sẽ gồm nhưng bước cơ bản như sau:

- Loại bỏ ký tự không cần thiết: Xóa bỏ các ký tự đặc biệt không phải là chữ và số như dấu ngắt dòng, tab, những ký hiệu.
- Chuẩn hóa về chữ hóa/chữ thường: Chuyển đổi tất cả chữ cái trong chuỗi văn bản thành chữ thường (hoặc chữ hoa) đảm bảo nhất quán
- Tách từ và loại bỏ từ dừng (Stop Words): Tách từng từ ra khỏi câu và loại bỏ các từ dừng (ví dụ: "và", "là", "của") không mang nhiều ý nghĩa trong phân tích.
- Loại bỏ nhiễu và sửa lỗi chính tả: Xác định và sửa các lỗi chính tả, cũng như loại bỏ các phần nhiễu trong văn bản
- Tokenization: Chuyển đổi văn bản thành một chuỗi các token (có thể là từ, cụm từ, hoặc ký tự) để xử lý dễ dàng hơn.
- Vector hóa: Chuyển đổi văn bản thành dạng vector sử dụng các kỹ thuật như Bag of Words, TF-IDF, hoặc Word Embeddings để máy tính có thể xử lý.

Dữ liệu hình ảnh

Thu thập dữ liệu hình ảnh là một việc không còn xa lạ gì, dữ liệu có thể thu thập từ Internet, các trang mạng xã hội và nền tảng chia sẽ hình ảnh như Instagram, Facebook. Dữ liệu thu được nhìn chung thời điểm hiện tại sự phát triển của công nghệ và trang thiết bị chất lượng hình ảnh khá là tốt, tuy nhiên vẫn còn đó những hình ảnh chưa đạt được chất lượng ổn như ảnh bị mờ nhiễu, hình ảnh bị méo xệnh v.v. Chúng ta có thể sử dụng một số phương pháp để cải thiện chất lượng hình ảnh giảm nhiễu, tăng cường độ sáng và tương phản, cân bằng màu, phục hồi ảnh v.v. Ngoài ra, còn có thể sử dụng một số mô hình học sâu để cải thiện chất lượng hình ảnh.

Nhìn chung, công việc tiền xử lý dữ liệu văn bản và hình ảnh là công việc khá tốn thời gian. May mắn thay trên công đồng đã có khá nhiều bộ dữ liệu đã được xử lý như Bảng 2 .3, phần nào đã rút ngắn được quá trình này.

Nguồn	**Số lượng**
MSCOCO	600k image/text
Lainon5B	5B image/text
cc12m	12M image/text
ImageNet	14M image/text

Bảng 2.3: Tổng hợp số lượng các ảnh/văn bản từ các nguồn dữ liệu

2.2. Trích chọn đặc trưng

Sau khi có được dữ liệu thô là các cặp hình ảnh và mô tả văn bản kèm theo. Để máy tính có thể hiểu và xử lý, dữ liệu này sẽ được chuyển đổi thành các đặc trưng biểu diễn dưới dạng vector. Dữ liệu hình ảnh sẽ được mã hóa thông qua các mạng như ResNet, Vision Transformer. Còn dữ liệu văn bản sẽ là Text Transformer, CBOW.

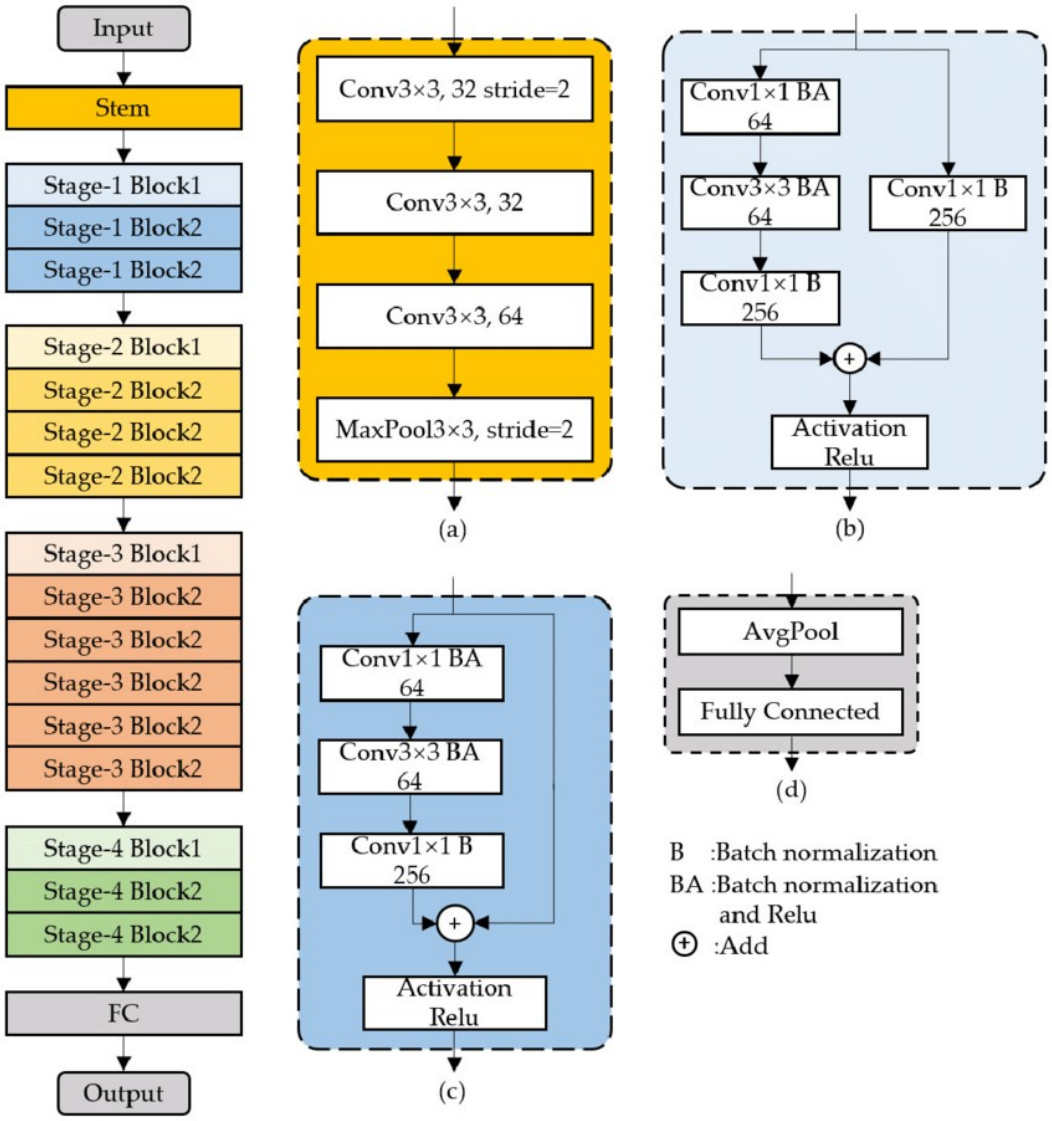
2.2.1. Image encoder

2.2.1.1. Mạng ResNet

Residual Network (ResNet) là một mô hình học sâu được sử dụng cho các ứng dụng thị giác máy tính. Đó là Convolutional Neural Network (CNN) được thiết kế để hỗ trợ hàng trăm hoặc hàng nghìn lớp tích chập. Các kiến trúc CNN trước đây không thể mở rộng quy mô thành số lượng lớn các lớp, dẫn đến hiệu suất bị hạn chế. Tuy nhiên, khi thêm nhiều lớp hơn, các nhà nghiên cứu phải đối mặt với vấn đề "độ dốc biến mất" [13]. Mạng lưới thần kinh được đào tạo thông qua quy trình lan truyền ngược dựa trên độ dốc giảm dần, dịch chuyển hàm mất mát xuống và tìm các trọng số giảm thiểu nó. Nếu có quá nhiều lớp, các phép nhân lặp đi lặp lại cuối cùng sẽ làm giảm độ dốc cho đến khi nó "biến mất" và hiệu suất sẽ bão hòa hoặc suy giảm sau mỗi lớp được thêm vào.

ResNet cung cấp một giải pháp sáng tạo cho vấn đề biến mất độ dốc, được gọi là "skip connections". ResNet sắp xếp nhiều ánh xạ nhận dạng (các lớp chập không làm gì lúc đầu), bỏ qua các lớp đó và sử dụng lại các kích hoạt của lớp trước đó. Việc bỏ qua sẽ tăng tốc quá trình đào tạo ban đầu bằng cách nén mạng thành ít lớp hơn. Sau đó, khi mạng được đào tạo lại, tất cả các lớp sẽ được mở rộng và các phần còn lại của

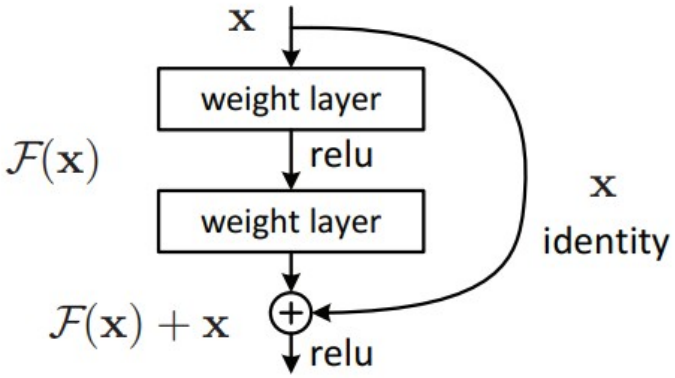
mạng được gọi là phần dư được phép khám phá thêm không gian đặc trưng của hình ảnh đầu vào như Hình 2 .3.



Hình 2.3: Kiến trúc mạng ResNet-50 (Nguồn: https://www.researchgate.net/).

Residual Block

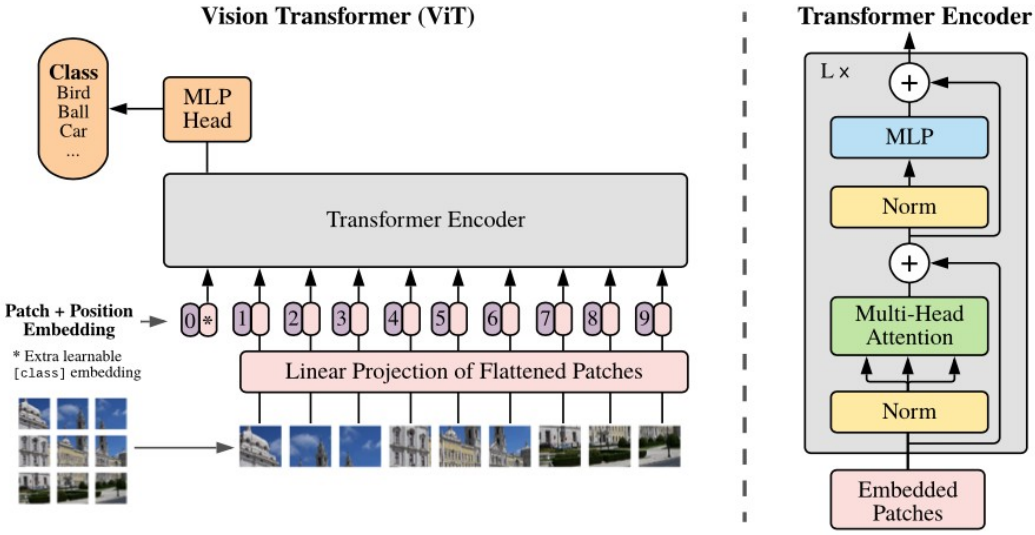
Residual block là một phần quan trọng của kiến trúc ResNet. Trong các kiến trúc cũ hơn như VGG16, các lớp tích chập được xếp chồng lên nhau với các lớp chuẩn hóa hàng loạt và kích hoạt phi tuyến như ReLu ở giữa chúng. Phương pháp này hoạt động với một số lượng nhỏ lớp chập mức tối đa cho mô hình VGG là khoảng 19 lớp. Tuy nhiên, nghiên cứu tiếp theo phát hiện ra rằng việc tăng số lớp có thể cải thiện đáng kể hiệu suất của CNN. Kiến trúc ResNet giới thiệu khái niệm đơn giản về việc thêm đầu vào trung gian vào đầu ra của một chuỗi khối tích chập.



Hình 2.4: Tổng quan về Residual block điển hình (Nguồn: https://www.researchgate.net/)

Hình 2 .4 trên cho thấy một khối dư (Residual block) điển hình. Điều này có thể được biểu thị bằng cách sử dụng biểu thức đầu ra bằng F(x) + x trong đó x là đầu vào cho residual block và đầu ra từ lớp trước và F(x) là một phần của CNN bao gồm một số khối chập.

2.2.1.2. Vision Transformer



Hình 2.5: Kiến trúc cơ bản về Vision Transformer (Nguồn: https://www.researchgate.net/)

Vision Transformer (ViT) là một kiến trúc mô hình học sâu đột phá được giới thiệu bởi Google Research vào cuối năm 2020. ViT áp dụng kiến trúc Transformer, trước đó chủ yếu được sử dụng trong xử lý ngôn ngữ tự nhiên (NLP), cho lĩnh vực thị giác máy

tính. Vision Transformers là kiến trúc sử dụng cơ chế self-attention để xử lý hình ảnh. Kiến trúc Vision Transformer bao gồm một loạt các khối máy biến áp (Transformer). Mỗi khối máy biến áp bao gồm hai lớp con: một lớp multi-head self-attention và một lớp feed-forward như Hình 2 .5.

Như Hình 2 .5 lớp self-attention tính toán trọng số "attention" cho từng pixel trong hình ảnh dựa trên mối quan hệ của nó với tất cả các pixel khác, trong khi lớp feed-forward nguồn cấp dữ liệu áp dụng phép biến đổi phi tuyến tính cho đầu ra của lớp self-attention. Multi-head attention có vai trò mở rộng cơ chế này bằng cách cho phép mô hình tham dự đồng thời vào các phần khác nhau của chuỗi đầu vào. Các thành phần trong kiến trúc ViT cụ thể như sau:

- **Hình ảnh đầu vào**: Hình ảnh được nhập vào mô hình và chia thành nhiều phân đoạn (patches). Mỗi patch sau đó được làm phẳng và ánh xạ vào không gian vectơ thông qua một phép chiếu tuyến tính (Linear Projection of Flattened Patches).

- **Patch Embedding**: Các vectơ thu được từ bước trên được gọi là các patch embeddings. Chúng được sử dụng làm đầu vào cho bộ mã hóa Transformer.

- **Positional Embedding**: Vì Transformer không có khả năng tự nhiên để nhận biết thông tin vị trí không gian, thông tin vị trí (positional embedding) được thêm vào cho mỗi patch embedding để bảo tồn thông tin vị trí không gian của chúng trong hình ảnh.

- **Transformer Encoder**: Đây là trái tim của ViT. Bộ mã hóa Transformer bao gồm một chuỗi các khối lặp lại, mỗi khối bao gồm hai thành phần chính:

  Multi-Head Attention: Cho phép mô hình chú ý đến các phần khác nhau của hình ảnh đồng thời, giúp nó học được mối quan hệ giữa các phân đoạn.

  Feed-Forward Neural Network (MLP): Một mạng nơ-ron tiếp theo xử lý thông tin sau cơ chế attention để tiếp tục trích xuất các đặc trưng.

  Mỗi khối cũng có các lớp chuẩn hóa (Norm) và cộng dư (residual connections) giúp cải thiện quá trình học và ổn định mô hình.

- **MLP Head**: Sau bộ mã hóa Transformer, một mạng nơ-ron tiếp theo (thường là một hoặc nhiều lớp kết nối đầy đủ) được sử dụng để phân loại. Đầu ra của lớp này chính là dự đoán phân loại cuối cùng cho hình ảnh.

- **Class Embedding**: Trong ViT, một embedding học được thêm vào đầu chuỗi các patch embeddings, đại diện cho toàn bộ hình ảnh. Sau quá trình mã hóa, đầu ra của phần này được sử dụng cho việc phân loại.

2.2.2. Text encoder

2.2.2.1. Text Transformer

Text Transformer, thường chỉ gọi là Transformer trong ngữ cảnh của xử lý ngôn ngữ tự nhiên (NLP). Mô hình Transformer, một kiến trúc mạng nơ-ron tiên tiến, được giới thiệu bởi Google trong một bài báo năm 2017, đã mở ra một hướng mới trong việc chuyển đổi dữ liệu từ dạng này sang dạng khác. Ban đầu được phát triển để nâng cao hiệu suất trong việc dịch máy từ tiếng Anh sang tiếng Pháp, mô hình này nhanh chóng chứng minh khả năng vượt trội của mình bằng việc cắt giảm đáng kể thời gian huấn luyện so với các kiến trúc trước đó [6].

Tuy nhiên, không lâu sau đó, các nhà nghiên cứu nhận thấy rằng khả năng của mô hình Transformer vượt xa phạm vi ban đầu. Nó đã được áp dụng rộng rãi, từ việc sinh văn bản và hình ảnh cho đến việc cung cấp hướng dẫn cho robot, cho thấy khả năng áp dụng đa dạng của nó. Transformer còn đóng một vai trò quan trọng trong lĩnh vực AI đa phương thức, nơi nó chuyển đổi ngôn ngữ tự nhiên thành các hình ảnh hoặc hướng dẫn cho robot, nhờ khả năng mô hình hóa các mối quan hệ giữa các loại dữ liệu khác nhau.

Trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), Transformer hiện là công cụ chuẩn mực, vượt trội hơn hẳn so với các phương pháp cũ. Ngoài ra, các mô hình dựa trên Transformer đã được chứng minh là có khả năng học hiểu cấu trúc phân tử hóa học, dự đoán cấu trúc ba chiều của protein, và thậm chí phân tích dữ liệu y tế với độ chính xác cao.

Một trong những yếu tố then chốt của Transformer là cơ chế chú ý (attention), cho phép mô hình xác định và tập trung vào các phần quan trọng của dữ liệu đầu vào. Cơ chế này giúp làm sáng tỏ ngữ cảnh cho từng từ hoặc token, dù là trong văn bản, hình ảnh, cấu trúc protein, hay thậm chí là trong các mẫu âm thanh, qua đó cải thiện đáng kể khả năng hiểu và xử lý thông tin phức tạp.

Attention

Cơ chế attention trong mô hình Transformer là một phát minh đột phá trong lĩnh vực học sâu, đặc biệt là trong xử lý ngôn ngữ tự nhiên (NLP). Cơ chế này giúp mô hình tập trung vào những phần quan trọng của dữ liệu đầu vào khi thực hiện dự đoán hoặc sinh đầu ra.

Ở cấp độ cơ bản, attention cho phép mô hình đánh giá mức độ liên quan của mỗi phần trong dữ liệu đầu vào so với các phần khác. Trong bối cảnh NLP, điều này có nghĩa là mô hình có thể tập trung vào các từ cụ thể trong câu khi nó đang cố gắng dịch hoặc sinh ra một câu mới. Mô hình Transformer sử dụng ba loại attention chính: self-attention, encoder-decoder attention, và cross-attention, mỗi loại phục vụ một mục đích khác nhau trong quá trình học [6].

- **Self-Attention**: Loại attention này cho phép mỗi vị trí trong chuỗi đầu vào tính toán mức độ tập trung của nó đối với mọi vị trí khác trong cùng một chuỗi. Điều này giúp mô hình cải thiện khả năng hiểu ngữ cảnh và mối quan hệ giữa các từ trong câu.

- **Encoder-Decoder Attention**: Trong giai đoạn giải mã, mô hình cần xem xét cả thông tin từ bộ mã hóa (encoder) và từ bộ giải mã (decoder) hiện tại. Cơ chế attention giúp decoder tập trung vào những phần quan trọng nhất của chuỗi đầu vào khi tạo ra mỗi từ tiếp theo.

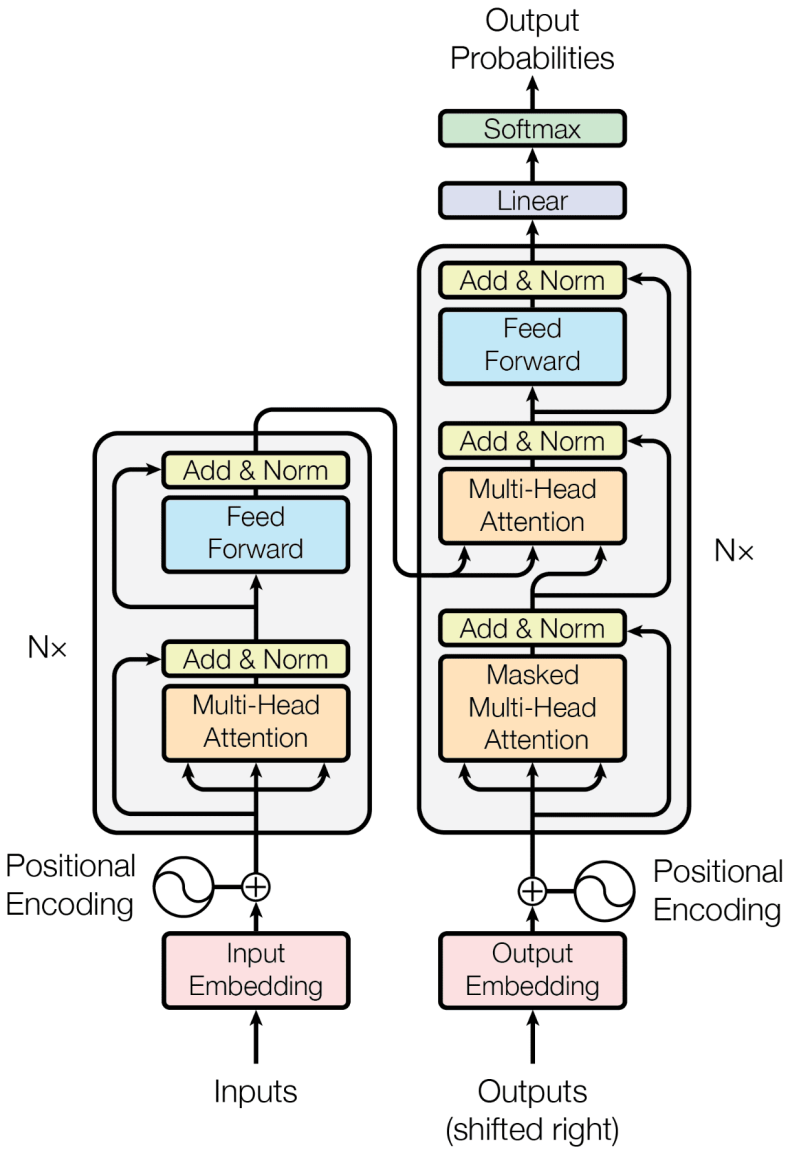
- **Cross-Attention**: Đây là một dạng attention nơi mà bộ giải mã tập trung vào đầu ra của bộ mã hóa. Nó thường được sử dụng trong các mô hình Transformer mới hơn và giúp cho việc truyền thông tin giữa hai phần của mô hình được linh hoạt hơn.

Cơ chế attention được định lượng thông qua việc tính điểm số attention, thường bằng cách sử dụng một hàm softmax để chuẩn hóa các điểm số này thành một phân phối xác suất. Những điểm số cao hơn tương ứng với mức độ chú ý cao hơn mà mô hình dành cho các phần của dữ liệu đầu vào khi thực hiện một nhiệm vụ cụ thể.

Kiến trúc mô hình

Kiến trúc transformer bao gồm bộ mã hóa (encoder) và bộ giải mã (decoder) hoạt động cùng nhau. Cơ chế attention cho phép transformer mã hóa ý nghĩa của từ dựa trên tầm quan trọng ước tính của các từ hoặc mã thông báo khác. Điều này cho phép các máy biến áp xử lý song song tất cả các từ hoặc mã thông báo để có hiệu suất nhanh

hơn, giúp thúc đẩy sự phát triển của các mô hình học sâu rất ngày càng lớn hơn. Kiến trúc transformer được thể hiện như Hình 2 .6:



Hình 2.6: Kiến trúc mạng Transformer [6].

- **Input Embedding**: Đầu vào của mô hình là các token từ văn bản được biến đổi thành vectơ nhúng. Mỗi token được ánh xạ sang một vectơ dày đặc qua một bảng nhúng.

- **Positional Encoding**: Mô hình Transformer không có khả năng tự nhiên để nhận biết thứ tự tuần tự của dữ liệu đầu vào, vì vậy positional encodings được thêm vào để cung cấp thông tin vị trí cho mỗi token. Điều này giúp mô hình biết được vị trí của các token trong chuỗi.

- **Multi-Head Attention**: Cơ chế này giúp mô hình tập trung vào những phần khác nhau của đầu vào khi đang xử lý một phần cụ thể. "Multi-head" có nghĩa là quá trình này được thực hiện nhiều lần với các trọng số khác nhau để mô hình có thể học cách tập trung vào nhiều khía cạnh khác nhau của dữ liệu.

- **Masked Multi-Head Attention**: Đặc biệt trong bộ giải mã, masked attention đảm bảo rằng dự đoán cho một vị trí chỉ có thể phụ thuộc vào các vị trí trước đó, giúp ngăn chặn việc "nhìn" vào tương lai trong quá trình dự đoán chuỗi tiếp theo.

- **Add & Norm**: Sau mỗi sub-layer, như Multi-Head Attention hoặc Feed Forward, kết quả được cộng với đầu vào ban đầu của sub-layer đó (cộng dư, residual connection), sau đó được chuẩn hóa. Điều này giúp tránh vấn đề biến mất gradient và giúp mô hình học được những thông tin tổng hợp từ các lớp trước.

- **Feed Forward**: Mỗi lớp trong encoder và decoder chứa một mạng feed forward độc lập, giúp xử lý thông tin từ attention layer.

- **Linear**: Lớp tuyến tính chuyển đổi từ không gian vectơ nhúng sang không gian có kích thước mong muốn cho đầu ra.

- **Softmax**: Lớp softmax được sử dụng để chuyển đổi các giá trị đầu ra của lớp tuyến tính thành một phân phối xác suất, cho biết khả năng của mỗi từ tiếp theo trong chuỗi.

2.3. Biểu diễn hình ảnh và văn bản trong không gian tiềm ẩn

Không gian ẩn là một không gian trừu tượng nhiều chiều, nơi các đặc trưng của dữ liệu được biểu diễn trong một dạng nén và tinh gọn. Điểm nổi bật của không gian này là nó thường không trực tiếp quan sát được và được học thông qua các mô hình học máy. Mục đích chính của không gian ẩn là để phát hiện và biểu diễn các đặc trưng quan trọng và tiềm ẩn của dữ liệu, giúp trong việc giảm kích thước dữ liệu và làm nổi bật những thông tin có giá trị.

Trong không gian ẩn, dữ liệu thường được biểu diễn dưới dạng vect, nơi mỗi chiều của vectơ biểu diễn một đặc trưng tiềm ẩn của dữ liệu. Mô hình học máy học cách ánh

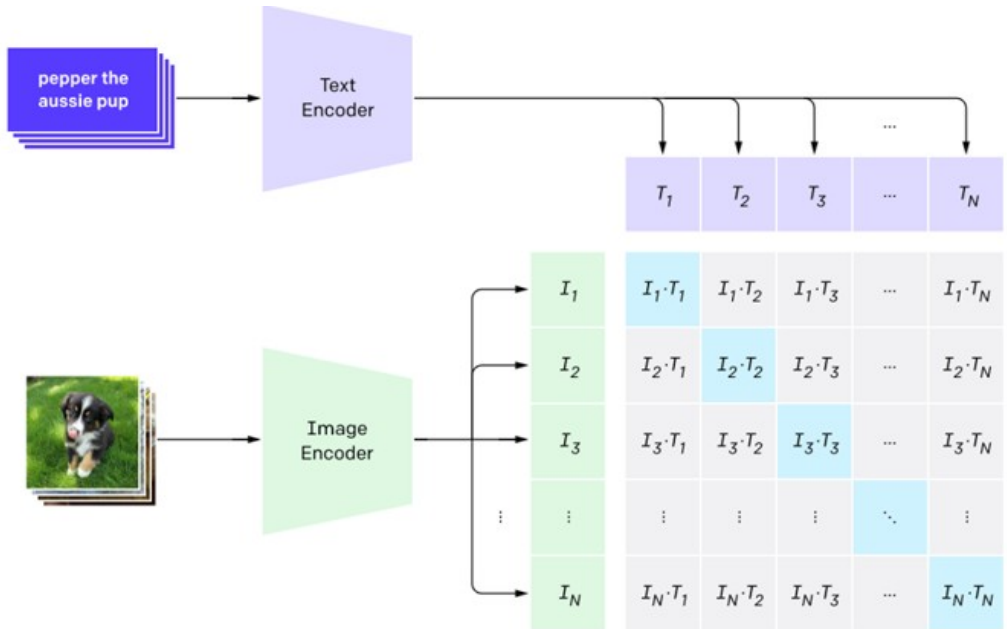
xạ dữ liệu từ không gian đầu vào (ví dụ: hình ảnh, văn bản) sang không gian ẩn và ngược lại. Quá trình này giúp tìm ra các mối quan hệ phức tạp và tiềm ẩn trong dữ liệu.

CLIP (Contrastive Language-Image Pre-Training)

Các hệ thống thị giác máy tính truyền thống được huấn luyện với một tập hợp cố định các danh mục đối tượng được xác định trước. Điều này hạn chế tính linh hoạt của chúng: mỗi khi chúng ta gặp một khái niệm trực quan mới, chúng ta cần đào tạo lại mô hình bằng các ví dụ được gắn nhãn về khái niệm này. CLIP sử dụng phương pháp học sâu để học cách tương quan giữa hình ảnh và văn bản. Nó được huấn luyện với lượng lớn dữ liệu bao gồm cặp hình ảnh và văn bản tương ứng, giúp nó phát triển khả năng liên kết sâu sắc giữa hai loại dữ liệu này. Ý tưởng chính của CLIP là huấn luyện bộ mã hóa hình ảnh và văn bản để tối đa hóa độ tương tự cosine của N cặp hợp lệ trong mỗi lô (và giảm thiểu độ tương tự của các cặp không hợp lệ).

CLIP đã được chứng minh có khả năng học hỏi các biểu diễn hình ảnh đầy mạnh mẽ, bao gồm cả yếu tố ngữ nghĩa và phong cách. Các nghiên cứu cho thấy rằng việc tạo ra các biểu diễn hình ảnh một cách minh bạch có thể nâng cao sự đa dạng trong hình ảnh mà vẫn giữ được độ chân thực và mức độ tương đồng với chú thích. Thêm vào đó, không gian nhúng chung của CLIP cung cấp khả năng chỉnh sửa hình ảnh dựa trên văn bản mà không cần quá trình huấn luyện trước.

Kiến trúc mô hình



Hình 2.7: Quá trình huấn luyện tương phản của CLIP [14].

CLIP (Contrastive Language-Image Pre-Training) được hợp thành bởi 3 thành phần chính như Hình 2 .7. Một mạng nơ-ron tích chập được sử dụng để xử lý và biểu diễn hình ảnh, một mạng nơ-ron dựa trên transformer được sử dụng để xử lý và biểu diễn văn bản. Cả hai mạng này được huấn luyện để ánh xạ dữ liệu hình ảnh và văn bản vào cùng một không gian đặc trưng [15].

Triển khai của CLIP

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Hình 2.8: Triển khai Contrastive Language-Image Pre-Training [14].

Trong quá trình khởi tạo bộ mã hóa cho mô hình học sâu đa chế độ, chúng ta sử dụng hai loại bộ mã hóa khác nhau: bộ mã hóa hình ảnh và bộ mã hóa văn bản. Bộ mã hóa hình ảnh có thể là một mạng ResNet hoặc Vision Transformer, dùng để chuyển hình ảnh thành các đặc trưng hình ảnh (I_f). Bộ mã hóa văn bản, có thể là mạng CBOW hoặc Transformer, chuyển văn bản thành các đặc trưng văn bản (T_f).

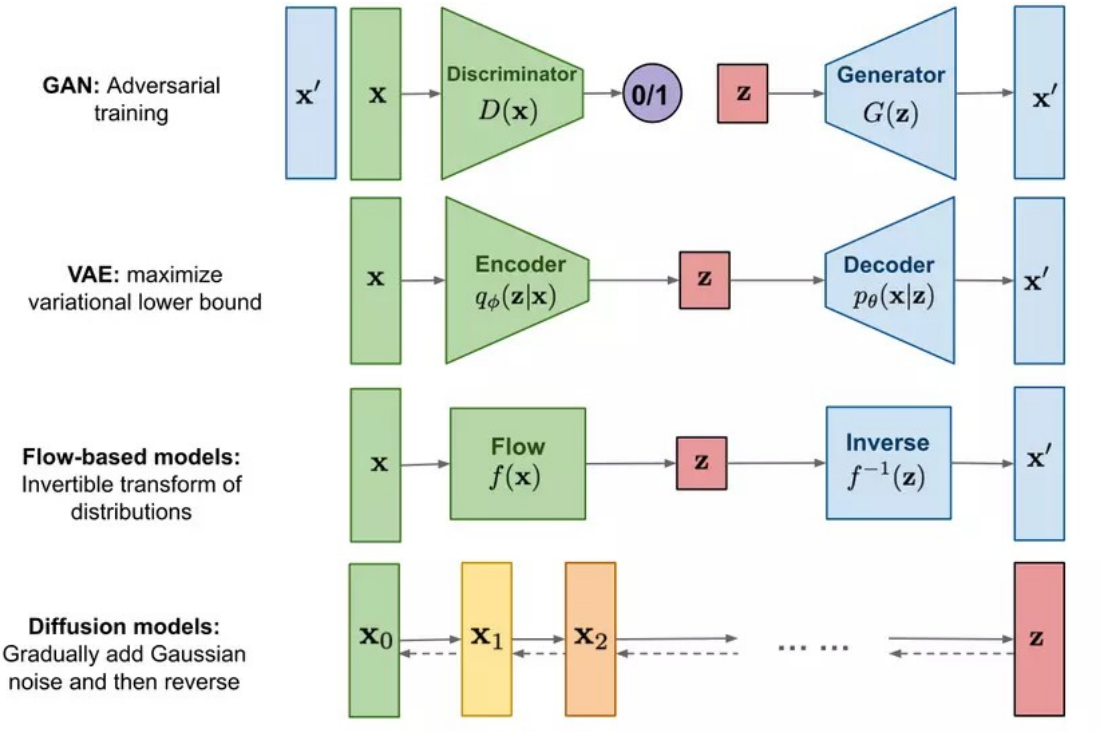
Sau đó, các đặc trưng này được tích hợp và chuẩn hóa để tạo ra biểu diễn đa chế độ (I_e và T_e) thông qua quá trình chuẩn hóa L2 và nhân với các trọng số đã học (W_i cho hình ảnh và W_t cho văn bản). Tiếp theo, mô hình tính toán độ tương đồng giữa các biểu diễn hình ảnh và văn bản thông qua độ tương đồng cosine, được điều chỉnh bằng tham số softmax temperature và biểu diễn dưới dạng logits. Cuối cùng, hàm mất

mát của mô hình được tính toán dựa trên mất mát cross-entropy cho cả hình ảnh và văn bản, sử dụng logits và nhãn đã định nghĩa. Mất mát tổng cộng là trung bình của mất mát từ hai chế độ này, đóng vai trò quan trọng trong việc huấn luyện mô hình để hiểu và kết hợp thông tin từ cả hình ảnh và văn bản.

2.4. Huấn luyện mô hình sinh cho phép tạo hình ảnh từ văn bản

2.4.1. Mô hình Diffusion

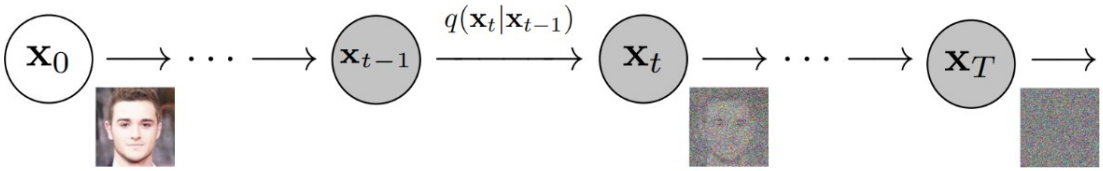
Những năm gần sự phát triển mạnh mẽ của các mô hình sinh với khả năng sinh hình ảnh từ đoạn văn bản, cho dù hình ảnh là không có thật hay phi logic. Hình 2 .9 là bốn loại mô hình nổi bật trong tác vụ sinh hình ảnh. Mô hình Diffusion là mô hình được nghiên cứu sau ô hình này đã khắc phục những điểm yếu, đồng thời thừa hưởng những điểm mạnh của mô hình trước đó như là một mô hình xác xuất có nhiệm vụ tạo ra một phân bổ cho các input tương tự như mô hình Flow-based, có một tầng được predetermined như GAN, và khả năng xấp xỉ phân phối dữ liệu được sinh ra với phân phối dữ liệu gốc tương tư như VAE. Nhờ tận dụng sức mạnh cốt lõi này đây được xem là mô hình tốt nhất trong khả năng sinh hình ảnh thời điểm hiện tại.



Hình 2.9: Các loại mô hình nổi bật trong mô hình sinh (Nguồn: https://pixta.vn/)

Mô hình khuếch tán (Diffusion), còn được gọi là mô hình xác suất khuếch tán hoặc mô hình tổng quát dựa trên điểm số, là một loại mô hình sinh (Generative model) [16]. Các mô hình khuếch tán được giới thiệu vào năm 2015 như một phương pháp để tìm hiểu một mô hình có thể lấy mẫu từ phân bố xác suất rất phức tạp. Họ sử dụng các kỹ thuật từ nhiệt động lực học không cân bằng, đặc biệt là khuếch tán. Ý tưởng chung là phá hủy một phân phối dữ liệu một cách từ từ và có kiểm soát thông qua một chuỗi cách khuếch tán thuận, việc của chúng ta lúc này là tìm cách học một mô hình có thể đảo ngược quá trình khuếch tán đó nhờ đó có thể phục hồi lại cấu trúc dữ liệu ban đầu [17].

Cụ thể hơn, mô hình khuếch tán là mô hình biến tiềm ẩn ánh xạ tới không gian tiềm ẩn bằng cách sử dụng chuỗi Markov cố định. Chuỗi này dần dần thêm nhiễu vào dữ liệu để thu được giá trị sau gần đúng $q(x_1|x_T)$, Ở đây $x_1,\ldots,x_T$ là các biến tiềm ẩn có cùng chiều với x_0. Trong Hình 2 .10 bên dưới, chúng ta thấy chuỗi Markov như vậy được biểu thị cho dữ liệu hình ảnh.



Hình 2.10: Quá trình minh họa thêm nhiễu vào hình ảnh (Nguồn: https://pixta.vn/)

Cuối cùng, hình ảnh được biến đổi tiệm cận thành nhiễu Gaussian thuần túy. Mục tiêu của việc đào tạo mô hình khuếch tán là tìm hiểu quy trình ngược lại, tức là đào tạo $P_0(x_{t-1}|x_t)$. Bằng cách di chuyển ngược lại dọc theo chuỗi này, chúng ta có thể tạo ra dữ liệu mới.

Mô hình Diffusion bao gồm 2 quá trình như sau: quá trình khuếch tán thuận (Foward diffusion process) và khuếch tán ngược (Reverse diffusion process).

Quá trình khuếch tán thuận

Tương tự như VAE, quá trình có thể được xem là quá trình encoder của mô hình diffusion. Ban đầu một hình ảnh được xem như là một điểm dữ liệu đầu vào ký hiệu là x_0 $q_{(x)}$ trong đó q(x) là phân phối của quá trình huấn luyện sẽ được thêm nhiễu từ từ theo từng bước (step). Khi mỗi điểm dữ liệu được thêm nhiễu nhữ vây, khi đó ta có thể biểu diễn phân phối của ảnh bị nhiễu tại hai thời điểm liên tiếp nhau (t -1, t) đó được cọi là quá trình thuận $q(x_t|x_{t-1})$ được tính như công thức (2.1) sau.

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \quad (2.1)$$

Như công thức (1) có thể thấy nhiễu được thêm vào từ phân phối Gauss bằng cách sample ra x_t với kỳ vọng $\sqrt{1-\beta_t}x_{t-1}$ và phương sai $\beta_t I$. Trong đó β ở đây là siêu tham số (hyperparrameter) được thay đổi theo thời gian có nhiệm vụ kiểm soát lượng nhiễu được thêm vào tại mỗi step t ở đó $\beta_t \in (0,1)$ hay còn được gọi lịch trình phương sai. Khi số step T đủ lớn, ảnh ban đầu sẽ chuyển thành một hình ảnh hoàn toàn nhiễu. Nếu mô hình hoạt động tốt khi đó x_T sẽ chính là gaussian tiêu chuẩn với phương sai là 1 và trung bình là 0 ký hiệu N (0, I).

Quá trình khuếch tán ngược

Mục tiêu của quá trình này là đảo ngược quá trình thuận thành $q(x_{t-1} \vee x_t)$ thì từ một hình ảnh bị nhiễu hoàn toàn vừa thu được từ quá trình trên chúng ta có thể đưa hình ảnh về trạng thai tương đương với ảnh đầu vào được lấy trong phân phối $q_{(x)}$ [8] [2]. Vì vậy chúng ta cần huấn luyệ mô hình p_θ có khả năng tương ứng với quá trình khuếch tán thuận. Được biểu diễn như công thức (2.2).

$$p(x_{t-1} \vee x_t) = N(x_{t-1}; \mu_{\theta(x_t,t)}, \Sigma_\theta(x_t,t)) \quad (2.2)$$

Trong đó $\mu_\theta$và Σ_θ là hai tham số mà ta cần ước lượng xấp xỉ. Nhờ vào tính chất của chuỗi Markov xác suất của mỗi sự kiện chỉ phụ thuộc vào trạng thái của sự kiện ngay trước đó từ đó quá trình biến đổi ngược từ x_T về x_0 được biểu diễn theo công thức (2.3).

$$P_0(x_{0:T}) = P(x_T)\prod_{t=1}^{T} P_\theta(x_t|x_{t-1}) \quad (2.3)$$

Như vậy mục tiêu lúc này là tối ưu hóa hàm negative log-likelihood (một hàm mất mát), tuy nhiên việc tính toán hàm này là khá khó khăn, chúng ta sẽ áp dụng một kỹ thuật là Variational lower bound [18] để có thể tính toán được hàm mất mất (loss funtion) từ đó ta có thể tính toán thông qua việc noise ϵ_t. Cuối cùng mô hình chỉ cần dựa đoán noise đã được thêm vào.

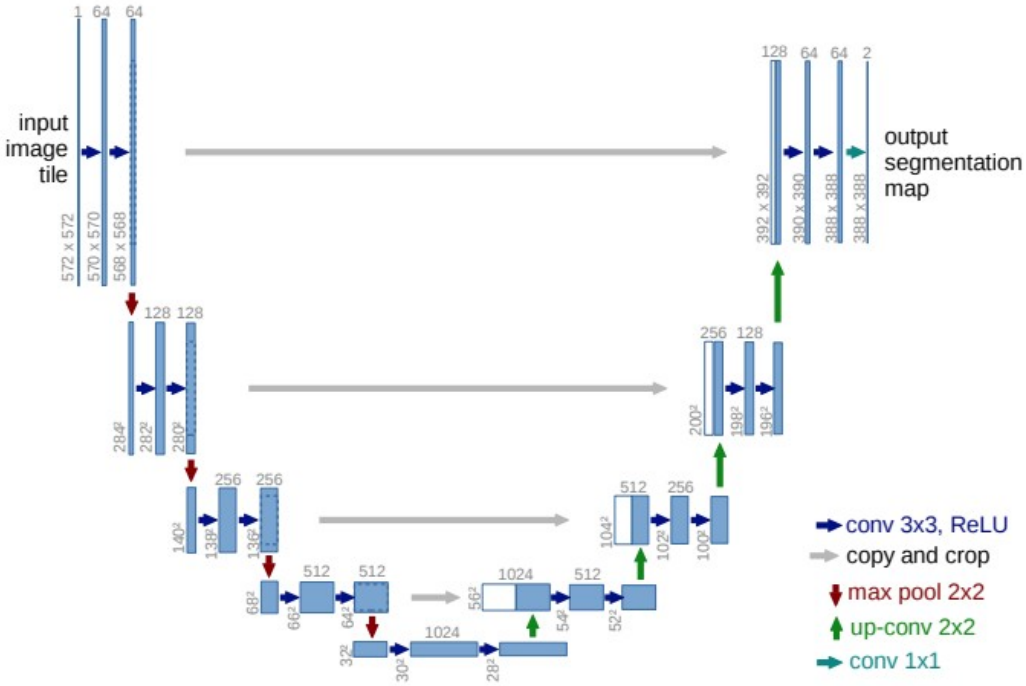
$$L_t^{simple} = E_{t\ [1,T],x_0,\epsilon_t}[\|\epsilon_t - \epsilon_\theta(x_t,t)\|^2] \quad (2.4)$$

Sau khi dự đoán được noise đã được thêm vào, để có hình ảnh tại các step chúng ta chỉ cần lấy hình ảnh nhiễu "trừ đi" noise. Vậy để dự đoán được phần noise chúng ta cần một mô hình có thể xác định được ngữ nghĩa của ảnh đầu vào từ đó xác định được

noise. Ngoài ra, kích thước đầu vào và đầu ra phải giống nhau. Do đó mạng Unet được đánh giá là có thể đáp ứng đầy đủ các yều đề ra ở trên.

Mạng Unet

Unet là một kiến trúc được phát triển bởi Olaf Ronneberger et al. Phân đoạn hình ảnh y sinh năm 2015 tại Đại học Freiburg, Đức. Đây là một trong những cách tiếp cận được sử dụng phổ biến nhất trong bất kỳ nhiệm vụ phân đoạn ngữ nghĩa nào hiện nay. Đó là một mạng lưới thần kinh tích chập hoàn toàn được thiết kế để học từ ít mẫu đào tạo hơn. Đó là một cải tiến so với FCN hiện có - "Mạng tích chập hoàn toàn để phân đoạn theo ngữ nghĩa" được phát triển bởi Jonathan Long và cộng sự. trong năm 2014) [17].



Hình 2.11: Kiến trúc mạng Unet (Nguồn: https://www.researchgate.net/)

Như Hình 2 .11 Unet là kiến trúc mạng bộ mã hóa-giải mã hình chữ U, bao gồm bốn khối bộ mã hóa và bốn khối bộ giải mã được kết nối qua một "cây cầu". Mạng bộ mã hóa có kích thước không gian bằng một nửa và gấp đôi số lượng bộ lọc (kênh đặc trưng) tại mỗi khối bộ mã hóa. Tương tự, mạng bộ giải mã tăng gấp đôi kích thước không gian và một nửa số kênh đặc trưng. Các thành phần chính của mạng Unet được thể hiển như sau.

- **Encoder**

Mạng mã hóa hoạt động như một bộ trích xuất đặc trưng và học cách biểu diễn trừu tượng của hình ảnh đầu vào thông qua một chuỗi các khối bộ mã hóa. Mỗi khối bộ mã hóa bao gồm hai tích chập 3x3, trong đó mỗi tích chập được theo sau bởi một hàm kích hoạt ReLU (Đơn vị tuyến tính chỉnh lưu). Chức năng kích hoạt ReLU đưa tính phi tuyến tính vào mạng, giúp khái quát hóa dữ liệu huấn luyện tốt hơn. Đầu ra của ReLU hoạt động như một kết nối bỏ qua cho khối giải mã tương ứng.

Ngoài ra, việc tuân theo phép gộp tối đa 2x2, trong đó kích thước chiều (chiều cao và chiều rộng) của bản đồ đặc trưng (feature maps) giảm đi một nửa. Điều này làm giảm chi phí tính toán bằng cách giảm số lượng tham số có thể huấn luyện.

- **Skip connection**

Các kết nối bỏ qua này cung cấp thông tin bổ sung giúp bộ giải mã tạo ra các đặc điểm ngữ nghĩa tốt hơn. Chúng cũng hoạt động như một kết nối phím tắt giúp truyền gradient gián tiếp đến các lớp trước đó mà không bị suy giảm. Nói một cách đơn giản, chúng ta có thể nói rằng việc bỏ qua kết nối giúp tạo ra luồng gradient tốt hơn trong khi lan truyền ngược, từ đó giúp mạng học cách biểu diễn tốt hơn.

- **Bridge**

Bridge kết nối bộ mã hóa và mạng bộ giải mã và hoàn thành luồng thông tin. Nó bao gồm hai tích chập 3x3, trong đó mỗi tích chập được theo sau bởi một hàm kích hoạt ReLU.

- **Decoder**

Mạng giải mã được sử dụng để lấy biểu diễn trừu tượng và tạo mặt nạ phân đoạn ngữ nghĩa. Khối giải mã bắt đầu bằng tích chập chuyển vị 2x2. Tiếp theo, nó được nối với feature maps bỏ qua kết nối tương ứng từ khối bộ mã hóa. Các kết nối bỏ qua này cung cấp các tính năng từ các lớp trước đó đôi khi bị mất do độ sâu của mạng. Sau đó, hai tích chập 3x3 được sử dụng, trong đó mỗi tích chập được theo sau bởi hàm kích hoạt ReLU. Đầu ra của bộ giải mã cuối cùng đi qua tích chập 1x1 với kích hoạt sigmoid. Hàm kích hoạt sigmoid cung cấp mặt nạ phân đoạn thể hiện sự phân loại theo pixel.

2.4.2. GLIDE

Như trình bày ở trên thì mô hình diffusion sau khi được huấn luyện có thể thực hiện sinh hình ảnh mới tương đồng với ảnh có trong phân phối dùng để huấn luyện từ một ảnh nhiễu hoàn toàn bất kỳ, tuy nhiên ảnh được sinh ra hoàn toàn ngẫu nhiễn, chúng ta

không thể biết được các đặc điểm cụ thể có trong ảnh, ngoài ra chất lượng hình ảnh được sinh ra thường có độ phân giải thấp. Để có thể sử dụng mô hình diffusion cho việc sinh ảnh theo mô tả văn bản thì cần cải tiến cũng như kết hợp các kỹ thuật để có thể cho ra hình ảnh như ý.

Classifier-free guidance

Kỹ thuật classifier-free guidance [9] trong quá trình sinh hình ảnh, giúp tăng cường sự phù hợp của hình ảnh sinh ra với mô tả văn bản mà không cần đến bộ phân loại riêng biệt. Đối với hướng dẫn không cần phân loại, nhãn y (văn bản) trong mô hình khuếch tán có điều kiện lớp $\varepsilon_\theta(x_t|y)$ được thay thế bằng nhãn null $\emptyset$ với xác suất cố định trong quá trình đào tạo. Trong quá trình lấy mẫu, đầu ra của mô hình được ngoại suy thêm theo hướng $\varepsilon_\theta(x_t|y)$và $\varepsilon_\theta(x_t|\varnothing)$ được biểu diễn như công thức (2.5).

$$\hat{\varepsilon}_\theta(x_t|y)=\varepsilon_\theta(x_t|\varnothing)+s.(\varepsilon_\theta(x_t|y)-\varepsilon_\theta(x_t|\varnothing)) \quad (2.5)$$

Kỹ thuật này mang hai đặc điểm nổi bật. Thứ nhất, nó cho phép một mô hình sử dụng kiến thức của chính mình trong quá trình được hướng dẫn, thay vì phải dựa vào kiến thức từ một mô hình phân loại khác (thường nhỏ hơn). Thứ hai, kỹ thuật này làm cho việc chỉnh sửa thông tin khó dự đoán, như hình ảnh trở nên đơn giản hơn khi sử dụng bộ phân loại (có thể là văn bản).

GLIDE

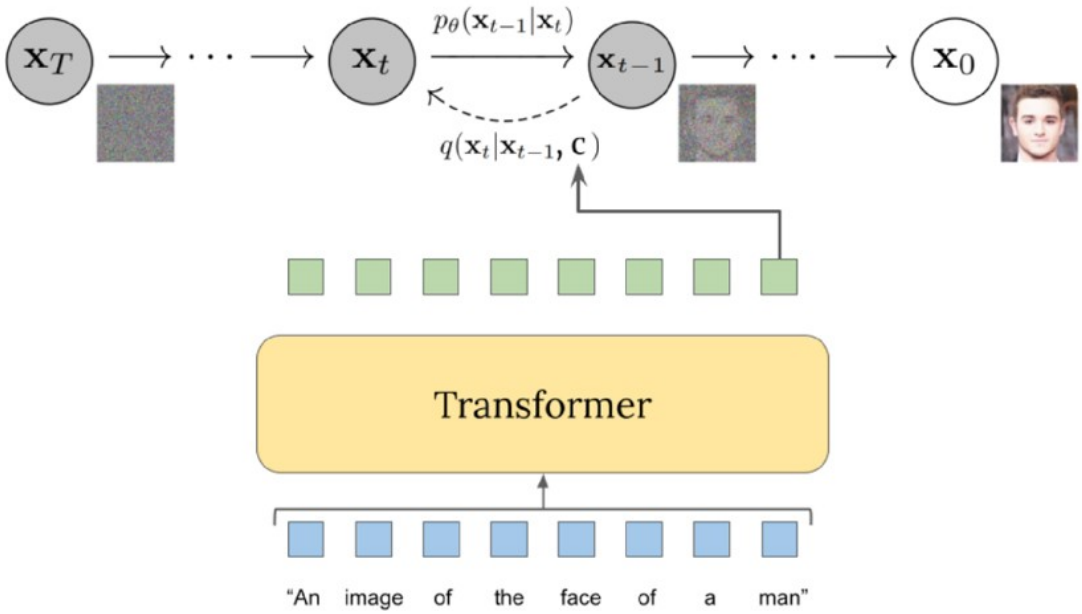
GLIDE (Guided Language to Image Diffusion for Generation and Editing) là một mô hình học máy phát triển bởi OpenAI, chuyên trong việc tạo ra hình ảnh từ các mô tả văn bản. Nó sử dụng kỹ thuật gọi là "diffusion model" kết hợp với hướng dẫn ngôn ngữ để tạo ra hình ảnh phức tạp và chi tiết từ mô tả văn bản. Do áp dụng kỹ thuật classifier-free guidance cho phép GLIDE cung cấp khả năng kiểm soát tinh vi và chỉnh sửa hình ảnh. Điều này có thể bao gồm việc điều chỉnh các chi tiết nhỏ, thay đổi màu sắc, hoặc thậm chí biến đổi các phần cụ thể của hình ảnh theo yêu cầu.

Điểm đặc biệt

Ở phần trước tại quá trình khuếch tán ngược của mô hình diffusion chúng ta cần dự đoán nhiễu được thêm vào từ đó phục hồi được hình ảnh về trạng thái trước đó. Vì vậy điểm đặc biệt của GLIDE so với mô hình diffusion là ở sẽ thêm thông tin về văn bản

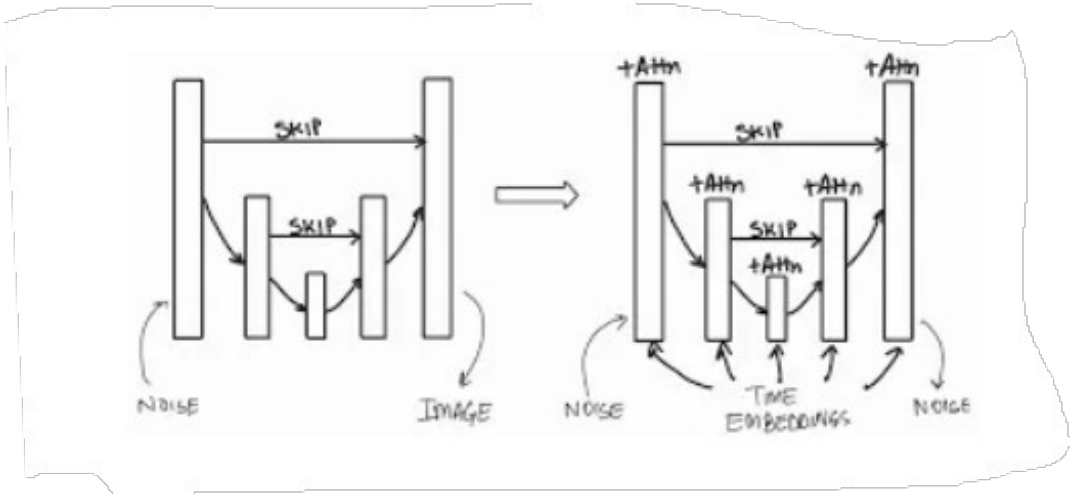
để hướng dẫn quá trình học của mô hình. Bởi vì đối với một mô hình diffusion đơn thuần, quá trình từ nhiễu trắng đến ảnh được sinh ra sẽ không có bất kì sự chỉ dẫn cụ thể nào để sinh ra một ảnh với một nội dung mong muốn cụ thể.

Phần văn bản này sẽ được mã hóa (encode) qua mô hình Transformer. Sau khi có được nhúng (embedding) từ Transformer, ta sẽ lấy embedding đó (c) đề làm điều kiện hướng dẫn cho mô hình diffusion. Được thể hiện thông qua Hình 2 .12.



Hình 2.12: Quá trình thêm hướng dẫn văn bản vào quá trình khuếch tán ngược (Nguồn: https://www.assemblyai.com/)

Để tích hợp embedding c. Ở quá trình denosie (khử nhiễu), trong bước học ε_θ dùng mạng Unet để thực hiện dữ đoán nhiễu ε_t.



Hình 2.13: Thêm điều kiện văn bản đã được mã hóa vào mạng unet (Nguồn: https://viblo.asia/)

Như hình Hình 2 .13, giá trịnh embeeding c được mã hóa từ văn bản sẽ được ánh xạ tới các lớp trung gian của mạng Unet thông qua cơ chế đó là cross-attention được biểu diễn thông qua công thức (2.6). Nhờ thêm vào quá trình này, mô hình diffusion từ đó cuối cùng cũng đã sinh ra hình ảnh mạng nội dung mong muốn từ mô tả văn bản.

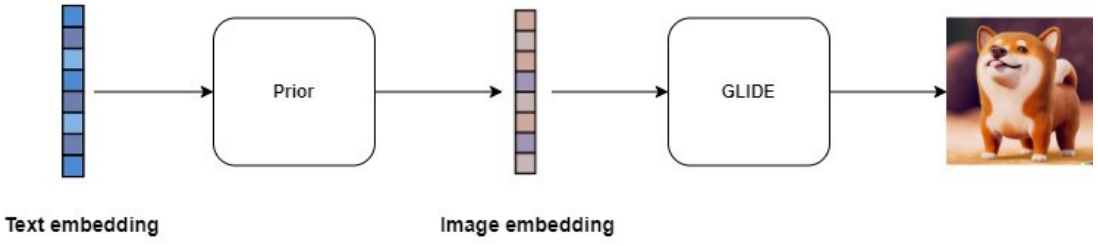
$$Attention(Q, K, V) = softmax\left(\frac{QK}{\sqrt{d}}V\right) \quad (2.6)$$

2.4.3. Prior

Mặc dù GLIDE có vai trò tạo ra hình ảnh phản ánh ngữ nghĩa được nắm bắt bởi các mã hóa hình ảnh. Nhưng việc tìm ra những biểu diễn trong mã hóa này, cần một bộ mã hóa văn bản. Prior có nhiệm vụ hiểu và diễn giải mô tả văn bản. Nó tạo ra một biểu diễn nội dung (content representation) từ văn bản nhập vào. Prior không tạo ra hình ảnh trực tiếp, nhưng nó sinh ra một dạng biểu diễn trừu tượng, thường được biểu diễn dưới dạng một vector hoặc một tập hợp các vector, mà mô tả các đặc điểm chính cần có trong hình ảnh cuối cùng.

Prior có thể được cấu trúc dưới dạng một mô hình tự hồi quy (Autoregressive Model) hoặc mô hình khuếch tán (Diffusion Model). Mô hình sẽ tận dụng khả năng của CLIP như đã giới thiệu ở phần Biểu diễn hình ảnh và văn bản trong không gian tiềm ẩn trong việc học cách mã hóa cả văn bản và hình ảnh. CLIP được huấn luyện để hiểu mối quan hệ giữa văn bản và hình ảnh, và Prior sử dụng mã hóa văn bản từ CLIP như một phần của quá trình của mình.

Trong quá trình lấy mẫu, để cải thiện chất lượng việc lấy mẫu được tiến hành ngẫu nhiên bằng cách sử dụng hướng dẫn không có bộ phân loại (classifier-free guidance) sẽ loại bỏ 10% đến 20% thông tin văn bản. Ngoài ra, trong thời gian lấy mẫu 2 image embedding được tạo ra với Prior sẽ được so sánh và lựa chọn ra cái có mô tả đúng ngữ nghĩa văn bản hơn thông qua CLIP.



Hình 2.14: Mô hình sinh kết hợp từ 2 mô hình diffusion đặc biệt.

Như Hình 2 .14, Text embdding chính là văn bản được mã hóa thông qua mô hình CLIP được sử dụng làm đầu vào của Prior, trong giai đoạn mô hình sẽ loại bỏ một lượng thông tin của văn bản để kích hoạt classifier-free guidance như đã trình bày ở trên. Sau qua trình trên chúng ta sẽ thu được Image embedding. Image embedding này sẽ được giải mã (Decode) thông qua GLIDE từ đó cho ra hình ảnh cuối cùng.

Chương 3

Thực nghiệm và đánh giá

Ở chương 3, đồ án sẽ trình bày về việc sử dụng tập dữ liệu có sẵn, những phương pháp và độ đo để đánh giá bài toán, và cuối cùng là kết quả thực nghiệm thu được qua các phần sau:

- Thu thập và xử lý dữ liệu
- Phương pháp và độ do đánh giá
- Kết quả thực nghiệm
- Kết luận

3.1. Thu thập và xử lý dữ liệu

Đồ án trên ý tưởng xây dựng mô hình sinh ảnh từ văn bản. Với ảnh sinh ra là ảnh chất lượng và đa dạng, nội dung hình ảnh phải phản ảnh đúng như mô tả văn bản. Ngoài ra, ảnh sinh ra là hình ảnh mới dựa trên dữ liệu của ảnh đào tạo. Vì thế để đạt được yêu cầu như trên cần xây dựng trên một mô hình phức tạp với bộ dữ liệu sử dụng trong quá trình phải đủ lớn cũng như chất lượng hình ảnh và mô tả đi kèm với ảnh phải đạt chất lượng tốt. Để thu thập một bộ dữ liệu trên thực tế là khá tốn thời gian, chi phí. May mắn, đề tài sinh ảnh từ văn bản là một chủ đề hot môt số năm gần đây. Vì vậy tập dữ liệu từ cộng đồng khá phong phú. Các cơ sở dữ liệu cộng đồng có sẵn hình ảnh và mô tả văn bản tương ứng như ImageNet, COCO (Common Objects in Context), Visual Genome … Vậy nên trên phạm vi nghiên cứu của đồ án này, chúng ta sẽ sử dụng một tập dữ liệu có sẵn để sử dụng cho quá trình huấn luyện mô hình.

ImageNet

ImageNet là một cơ sở dữ liệu hình ảnh được tổ chức theo hệ thống phân cấp WordNet (hiện tại chỉ có các danh từ), trong đó mỗi nút của hệ thống phân cấp được mô tả bởi hàng trăm, hàng nghìn hình ảnh. ImageNet là một trong những tập dữ liệu lớn và rộng rãi nhất được sử dụng trong nghiên cứu thị giác máy tính, đặc biệt là trong lĩnh vực nhận dạng hình ảnh. Tập dữ liệu này được phát triển bởi các nhà nghiên cứu tại Đại học Stanford và các cộng sự từ nhiều trường đại học khác. Dữ liệu được cung cấp miễn phí cho các nhà nghiên cứu với mục đích sử dụng phi thương mại [19].

Quy mô và cấu trúc của tập dữ liệu này như Hình 3 .15:

- **Hình ảnh:** ImageNet chứa hơn 14 triệu hình ảnh được gắn nhãn.
- **Phân loại:** Hình ảnh được tổ chức theo hệ thống phân loại WordNet, nghĩa là mỗi bức hình được gắn với một hoặc nhiều từ ngữ (synsets) trong WordNet.
- **Số lượng danh mục:** Có hơn 20.000 danh mục, mỗi danh mục chứa nhiều hình ảnh khác nhau

Hình 3.15: Cấu trúc folder của ImageNet (Nguồn: https://livebook.manning.com)

ImageNet nổi tiếng nhờ vào cuộc thi ILSVRC, một thách thức hàng năm bắt đầu từ năm 2010 đến 2017, nơi các nhóm nghiên cứu cạnh tranh để phát triển các thuật toán nhận dạng hình ảnh với độ chính xác cao nhất. Cuộc thi này đã đóng góp lớn trong việc thúc đẩy tiến bộ của thuật toán học sâu, nhất là mạng nơ-ron tích chập (CNN), dẫn đến những cải tiến đáng kể trong hiệu suất nhận dạng hình ảnh.

Mặc dù đây là tập dữ liệu được đấy ra khá chất lượng nhưng vẫn sẽ gặp phải nhưn vấn đề như độ chính xác của nhãn mặc dù ImageNet được gắn nhãn bởi con người, nhưng không phải tất cả các nhãn đều hoàn toàn chính xác do sự phức tạp và đa dạng của thế giới thực. Hơn thế nữa, lo ngại về việc liệu tập dữ liệu này có đại diện đầy đủ cho sự đa dạng của hình ảnh trong thế giới thực hay không, đặc biệt là sau khi công nghệ và xã hội tiếp tục phát triển.

Tuy nhiên ImageNet vẫn tiếp tục là một nguồn dữ liệu cơ bản cho nghiên cứu và phát triển trong lĩnh vực trí tuệ nhân tạo, nhưng cộng đồng nghiên cứu cũng đang hướng tới việc tạo ra các tập dữ liệu mới và đa dạng hơn để phản ánh tốt hơn thế giới xung quanh.

3.2. Phương pháp và các độ đo đánh giá

Để đánh giá các bài toán sinh ảnh, chúng ta cần đánh gia trên nhiều yếu tố khác nhau. Khác với các bài toán như phân lớp (segmentation) hay là phát hiện đối tượng trong ảnh (object detection) chúng ta có thể đánh giá bằng những phương pháp thường dùng thì để đánh giá khả năng sinh ảnh của mô hình cần dựa vào 2 yếu tố chính sau:

- **Chất lượng ảnh**: Ảnh sinh ra phải có chất lượng cao, ngữ nghĩa trong hình ảnh phải đúng với mô tả văn bản hay còn gọi là tính trung thực của hình ảnh.

- **Độ đa dạng và phong phú:** Ảnh ở mỗi lần sinh phải đa dạng, cần sinh ra được nhiều ảnh khác nhau. Nếu mô hình mãi chỉ sinh ra được một vài ảnh hay thuộc cùng 1 lớp thì mô hình không khác như những bài toán tìm kiếm hình ảnh.

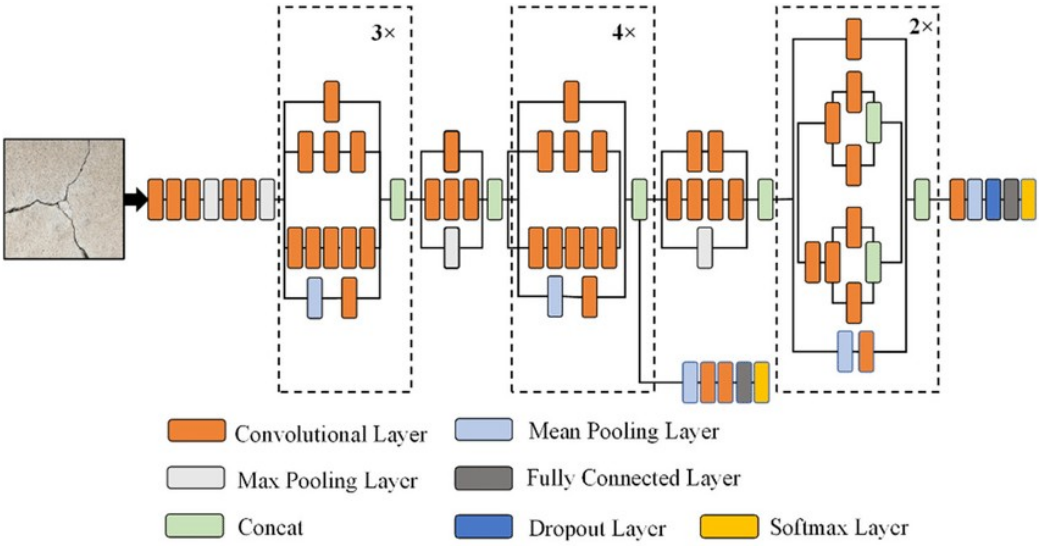
Vì vậy, để đánh giá mô hình sinh ảnh có điều kiện dựa trên mô tả văn bản có khá nhiều phương pháp khác nhau nhưng nhìn chung, sẽ được chia thành hai phương pháp chính đó là:

- **Đánh giá con người:** Dùng mắt thường để đánh giá, cách đánh giá thường sẽ dự đoán kết quả xem luôn. Tuy nhiên nó không khách quan (mỗi người có một định nghĩa về ảnh tốt đẹp khác nhau). Ngoài ra trên một tập dữ liệu lớn điều đó là không khả thi.

- **Đánh giá dựa trên chỉ số:** Việc sử dụng các chỉ số thỏa mãn 2 yếu tố trên và đồng thời áp dụng được cho tập dữ liệu lớn. Một số chỉ số hay được sử dụng như: Inception Score (IS), Frechet Inception Distance (FID), Image Quality Measures (SSIM, PSNR và Sharpness Difference) v.v.

Inception Score (IS)

Inception Score (IS) là một phương pháp đánh giá được sử dụng rộng rãi trong các bài toán sinh ảnh (image generation), đặc biệt là trong các mô hình sinh ảnh tự động như Generative Adversarial Networks (GANs), Diffusion Models. Giống như tên gọi Inception Score, trong đó thì inception chính là một một Convolutional Neural Network (CNN) dự đoán phân phối xác xuất của các lớp (class probabilities) cho mỗi hình ảnh, kiến trúc sử dụng Inception như Hình 3 .16.



Hình 3.16: Kiến trúc của mô hình Inception v3 (Nguồn: https://www.researchgate.net/)

Như Hình 3 .16 khi hình ảnh chạy qua các layer và cuối cùng sẽ qua lớp sofmax (là lớp cuối cùng) lúc này hình ảnh sẽ được phân phối xác xuất của các lớp, hình ảnh nếu rõ ràng thuộc một lớp nào đó sẽ xác suất cho ra tại lớp đó là rất cao, còn ảnh không thuộc lớp nào sẽ cho ra xác suất giữa các lớp là khá gần nhau (uniform). Từ đó dựa vào phân phối xác xuất mà phương pháp này đo lường được 2 yếu tố chính sau:

- **Kiểm tra được chất lượng ảnh**: Cho ảnh sinh ra thông qua mạng, nếu ảnh sinh ra sắc nét thì mạng sẽ có khả năng phân loại hình ảnh đó vào một lớp cụ thể (tức là xác suất của 1 lớp cao hơn hẳn).

- **Kiểm tra được độ đa dạng của ảnh sinh ra:** Tổng giá trị xác suất theo từng lớp của tất cả các ảnh sinh ra từ mô hình, nếu mô hình được xem ảnh sinh ra đa dạng thì tổng xác suất sẽ dạng uniform, ngược lại nếu dữ liệu chỉ sinh ra dữ liệu ở 1 hay 2 lớp thì tổng xác suất sẽ chỉ cao hơn ở 1 hay 2 lớp.

Để tính chỉ số này, Inception Socre sẽ được tính toán dựa trên sự phân kỳ Kullback-Leibler (KL divergence) giữa phân phối xác suất của từng ảnh và phân phối xác suất trung bình của toàn bộ tập dữ liệu sẽ được biểu diễn qua công thức (3.1).

$$IS = \exp ¿ \qquad (3.7)$$

Trong đó x là ảnh được sinh ra từ mô hình p(y|x) là phân phối xác suất dự đoán cho ảnh x bởi mạng Inception. p(y) sẽ là phân phối xác suất trung bình của các lớp trên toàn bộ tập ảnh sinh ra và D_{KL}là sự phân kỳ Kullback-Leibler giữa hai phân phối xác suất. Như vậy để ảnh sinh ra tốt (chất lượng tốt và đa dạng) thì chỉ số phải càng cao.

Tuy nhiên, hạn chế của Inception Score nếu mô hình chỉ sinh được một ảnh mỗi lớp thì chỉ số KL(Kullback–Leibler) hay chỉ số IS vẫn có thể cao, có nghĩa ảnh sinh ra chỉ đa dạng lớp nhưng không đa dạng ảnh trong mỗi lớp. Còn nếu mô hình sinh ra các ảnh trong dataset thì chỉ số KL cũng cao.

Frechet Inception Distance (FID)

FID đo lường sự khác biệt giữa hai tập hợp ảnh, thường là tập hợp ảnh thực và ảnh mô hình sinh ra thông qua cách tính toán sự khác biệt trong phân phối các đặc trưng của ảnh được trích xuất bởi một mô hình mạng nơ-ron sâu (thường là Inception v3) như Hình 3 .16. Và phân phối sử dụng sẽ là một phân phối Gauss nhiều chiều (dimension multivariate gaussian distribution).

Để các ảnh sinh ra giống các ảnh trong tập dữ liệu thì chúng ta mong muốn 2 phân phối Gauss nhiều chiều này sẽ giống nhau, hay giá trị trung bình (mean) và phương sai (variance) của 2 phân phối gần nhau và variance gần nhau được mô tả như công thức (3.2).

$$FID = \left\| \mu_{real} - \mu_{gen} \right\|^2 + Tr¿ \qquad (3.8)$$

Trong đó μ_{real}, μ_{gen}là giá trị trung bình của tập ảnh huấn luyện và tập ảnh được sinh ra, Σ_{real}, Σ_{gen} là ma trân hiệp phương sai tương ứng.

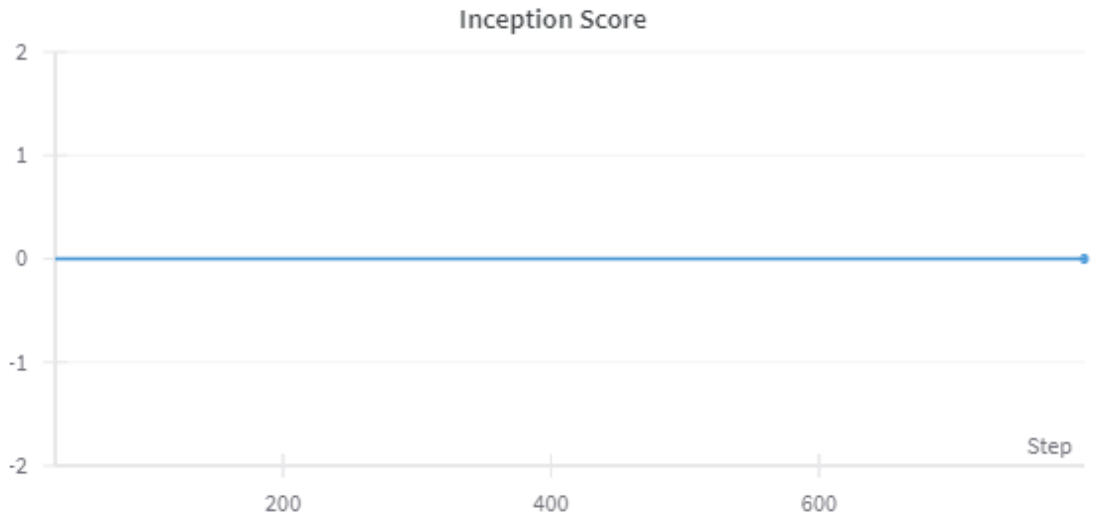
FID là một chỉ số quan trọng trong việc đánh giá chất lượng của ảnh sinh ra. Nó giúp so sánh mức độ giống nhau giữa tập ảnh thực và ảnh được sinh ra, qua đó đánh giá mức độ "thực tế" của các ảnh sinh. Tuy nhiên, giống như mọi chỉ số đánh giá, nó nên được sử dụng cùng với các phương pháp đánh giá khác để có cái nhìn toàn diện nhất.

3.3. Kết quả

Đồ án do phải huấn luyện một mô hình phức tạp, cần cấu hình thiết bị đám ứng được điều kiện để huấn luyện. Vì vậy ở phạm vi đồ án chỉ xây dựng mô hình ở quy mô nhỏ còn nhiều hạn chế ở kết quả thu được.

3.3.1. Kết quả với Inception Score

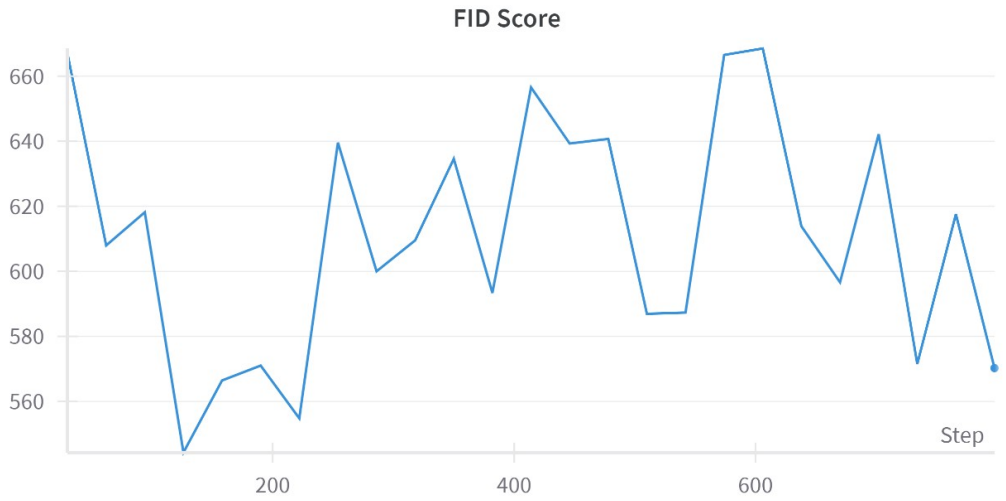
Biểu đồ Inception Score trên tập dự liệu ImageNet khi huấn luyện mô hình cho ra kết quả được thể hiện trong Hình 3 .17.



Hình 3.17: Biểu đồ Inception Score trên tập dữ liệu ImageNet

3.3.2. Kết quả với Frechet Inception Distance

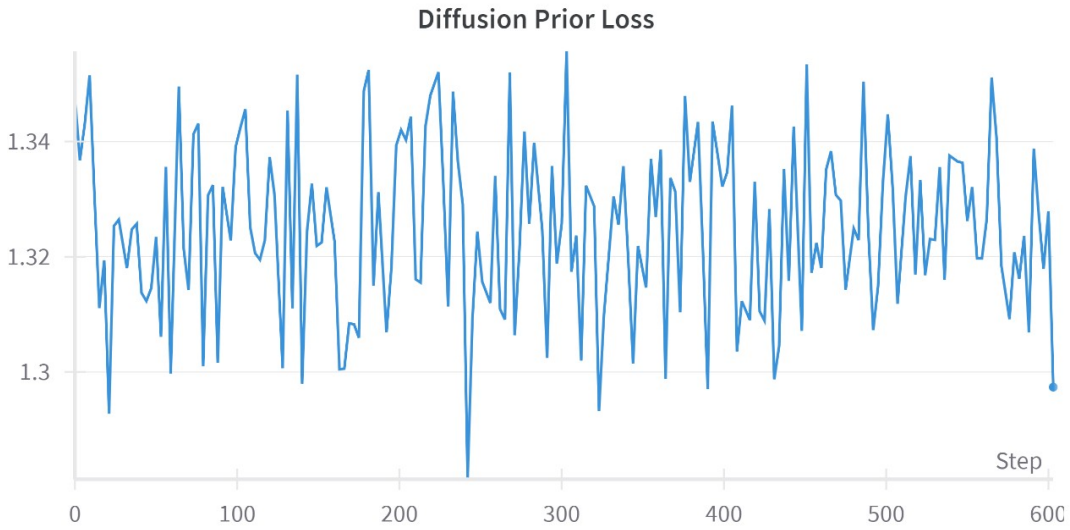
Biểu đồ Frechet Inception Distance trên tập dự liệu ImageNet khi huấn luyện mô hình cho ra kết quả được thể hiện trong Hình 3 .18.



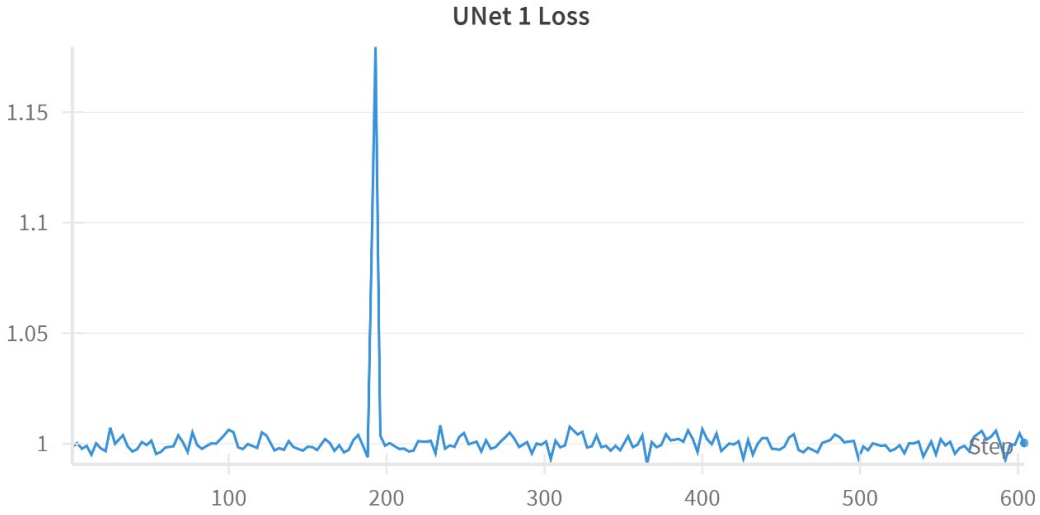
Hình 3.18: Biểu đồ Frechet Inception Distance trên tập dữ liệu ImageNet

3.3.3. Kết quả với Loss Function

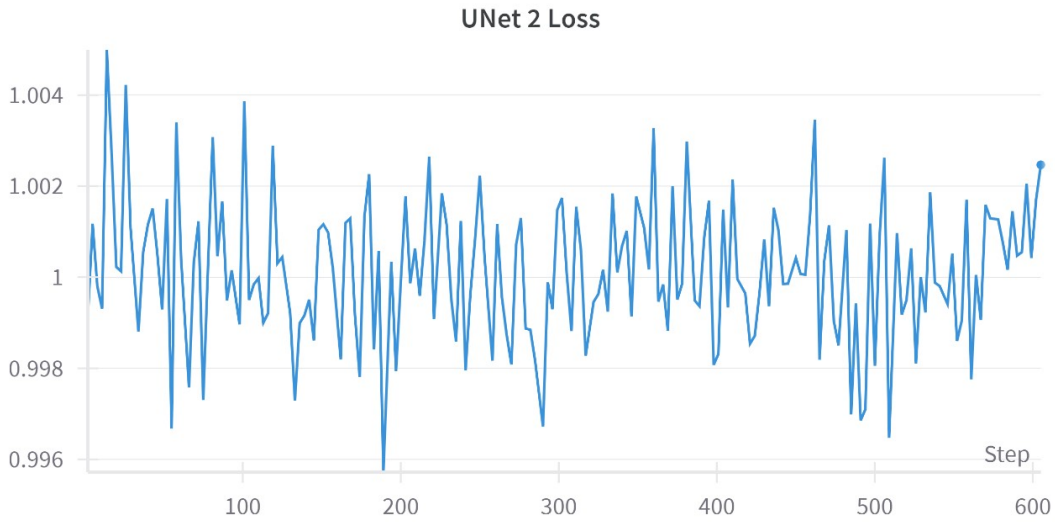
Biểu đồ chỉ số hàm mất mát (Loss function) tương ứng với 3 quá trình Diffusion prior, mạng Unet1 và mạng Unet2 khi huấn luyện trên tập dữ liệu train được thể hiện lần lượt ở Hình 3 .19, Hình 3 .20, Hình 3 .21.



Hình 3.19: Biểu đồ chỉ số mất mất của quá trình diffusion prior

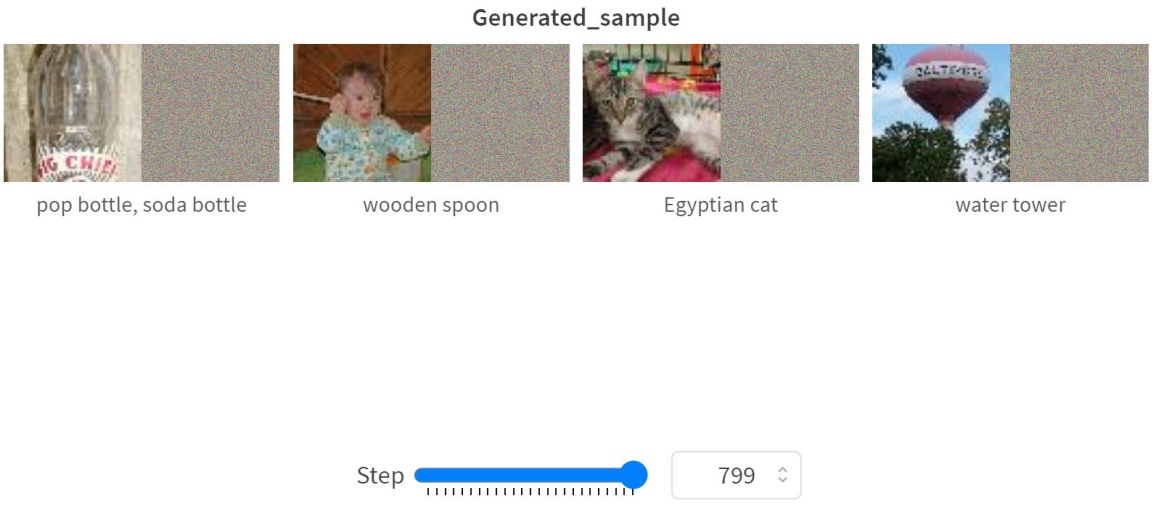


Hình 3.20: Biểu đồ chỉ số mất mất của mạng Unet1



Hình 3.21: Biểu đồ chỉ số mất mất của mạng Unet2

3.3.4. Kết quả mẫu thu được



Hình 3.22: Mẫu được tạo ra so với ảnh gốc

Như Hình 3 .22, các mẫu được tạo ra chỉ toàn nhiễu do mô hình chưa đáp ứng đủ thời gian huấn luyện mô hình. Điều này phản ánh đúng các chỉ số đánh giá IS rất thấp, và chỉ số FID rất cao được biểu diễn ở lần lượt các Hình 3 .17, Hình 3 .18.

Mạng	Số lượng parameters
Diffusion Prior	25M
Decoder	200M

Bảng 3.4: Số lượng trọng số ứng với các phần mô hình

Bảng 3 .4 thể hiện số lượng tham số của hai thành phần chính của mô hình đã đề xuất. Với 200 triệu tham số, Decoder là mô hình rất lớn và phức tạp. Cho thấy nó có khả năng xử lý và tạo ra dữ liệu phức tạp và chi tiết.

3.4. Kết luận

Với mục tiêu là tìm hiểu và phân tích và đánh giá nhưng phương pháp sinh hình ảnh từ mô tả hình ảnh. Đồ án đã đạt được một số kết quả sau:

- Giới thiệu tổng quan về trí tuệ nhân tạo sáng tạo, những loại dữ liệu, cũng như cấu trúc của chúng hay được dùng trong lĩnh vực này. Trình bày một số nghiên cứu nổi bật được sử dụng trong bài toán sinh hình ảnh từ mô tả văn bản bao gồm: hệ thống DM-GAN là hệ thống kết hợp giữa bộ nhớ động và GAN, hệ thống DALL-E.2 là hệ thống đang được đánh giá cao với việc sử dụng mô hình diffusion. Từ đó, giới thiệu khái quát những mô hình sinh là thành phần không thể thiếu trong bài toán sinh hình ảnh. Ngoài ra, nêu được mục tiêu và phương pháp sẽ tiếp cận trong đồ án.

- Trình bày tổng quan về mô hình sẽ tiếp cận, các phương pháp dùng để trích xuất những đặc trưng. Đối với hình ảnh sử mạng ResNet-50 và Vision Transfomer, với văn bản là Text Transformer. Sử dụng mô hình khuếch tán (diffusion models) kết hợp với một số kỹ thuật đặc biêt để nâng cao và kiểm soát chất lượng hình ảnh được sinh ra.

- Trình bày cấu trúc tập dữ liệu sử dụng xây dựng hệ thống. Giới thiệu tổng quan một số phương pháp và chỉ số dùng để đánh giá trong bài toán sinh hình ảnh.

- Kết quả thực nghiệm do chỉ xây dựng trên mô hình nhỏ, thời gian để huấn luyện mô hình hạn chế vì vậy hình ảnh sinh ra khá kém (bị nhiễu) tuy nhiên hướng đi mô hình khá tốt. Ngoài ra, chỉ số mất mát của các mạng khá thấp giao động 1%.

Hướng phát triển trong tương lai

- Mặc dù kết quả cho ra hình ảnh chất lượng khá kém do thời gian huấn luyện chưa đủ. Vì vậy cần huấn luyện mô hình với thời gian đủ dài. Ngoài ra, sử dụng bộ dữ liệu tốt hơn, với chất lượng hình ảnh tốt hơn, mô tả văn bản sát với những trường hợp thực tế hơn.

- Tối ưu mô hình, để giảm chi phí huấn luyện mà vẫn đảm bảo được chất lượng đem lại. Phát triển thêm các tính năng mới dựa trên mô hình hiện tại như chỉnh sửa hình ảnh, cá nhân hóa hình ảnh.

- Xây dựng hệ thống hoàn chỉnh đa ngôn ngữ, khả năng tích nhúng vào các hệ thống khác để từ đó có khả năng thương mại hóa.

TÀI LIỆU THAM KHẢO

[1] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, Patrick Zschech, "Generative AI," *arXiv:2309.07930,* 2023.

[2] Chenqiu Zhao, Guanfang Dong, Anup Basu, "Is Deep Learning Network Necessary for Image Generation?," *arXiv:2308.13612,* 2023.

[3] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang, "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis," *arXiv:1904.01310,* 2019.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, "Generative Adversarial Nets," *arXiv:1406.2661,* 2014.

[5] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv:2204.06125,* 2022.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need," *arXiv:1706.03762,* 2017.

[7] Matthias Freiberger, Peter Kun, Anders Sundnes Løvlie, Sebastian Risi, "CLIPMasterPrints: Fooling Contrastive Language-Image Pre-training Using Latent Variable Evolution," *arXiv:2307.03798,* 2023.

[8] Jonathan Ho, Ajay Jain, Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *arXiv:2006.11239,* 2020.

[9] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, Mark Chen, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," *arXiv:2112.10741,* 2021.

[10] Diederik P Kingma, Max Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114,* 2013.

[11] Gilad Cohen, Raja Giryes, "Generative Adversarial Networks," *arXiv:2203.00667,* 2022.

[12] Xingjian Zhen, Rudrasis Chakraborty, Liu Yang, Vikas Singh, "Flow-based Generative Models for Learning Manifold to Manifold Mappings," *arXiv:2012.10013,* 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385,* 2015.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv:2103.00020,* 2021.

[15] Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, Gao Huang, "Contrastive Language-

Image Pre-Training with Knowledge Graphs," *arXiv:2210.08901v1,* 2022.

[16] Calvin Luo, "Understanding Diffusion Models: A Unified Perspective," *arXiv:2208.11970,* 2022.

[17] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597,* 2015.

[18] Yuki Yasuda, Ryo Onishi, "A Theory of Evidence Lower Bound and Its Application to Super-Resolution Data Assimilation (SRDA) Using Conditional Variational Autoencoders," *arXiv:2308.03351,* 2023.

[19] Stanford Vision Lab, Stanford University, Princeton University, "ImageNet," 2020. [Online]. Available: https://image-net.org/.