**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KỲ**

**THỊ GIÁC MÁY TÍNH**

**Keras VQ-VAE for image generation**

*Người thực hiện*: **NGÔ HOÀNG NHẬT HUY - 20079071**

**NGUYỄN XUÂN GIANG - 20079601**

*Lớp:* **420301411801**

*Khoá:* **16**

*Người hướng dẫn*: **Thầy LƯU GIANG NAM**

**Thầy TRẦN TẤN THÀNH**

**TP HỒ CHÍ MINH, NĂM 2024**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KỲ**

**THỊ GIÁC MÁY TÍNH**

**Keras VQ-VAE for image generation**

*Người thực hiện*: **NGÔ HOÀNG NHẬT HUY - 20079071**

**NGUYỄN XUÂN GIANG - 20079601**

*Lớp:* **420301411801**

*Khoá:* **16**

*Người hướng dẫn*: **Thầy LƯU GIANG NAM**

**Thầy TRẦN TẤN THÀNH**

**TP HỒ CHÍ MINH, NĂM 2024**

# LỜI CẢM ƠN

Chúng em xin bày tỏ lòng biết ơn đến sự hướng dẫn, chỉ dẫn, và hỗ trợ của thầy Lưu Giang Nam và thầy Trần Tấn Thành đã cung cấp trong suốt quá trình thực hiện đồ án này. Những lời khuyên, ý kiến, và kiến thức mà thầy đã chia sẻ đã giúp chúng em hoàn thành công việc một cách hiệu quả hơn.

Chúng em cũng muốn bày tỏ lòng biết ơn đến tất cả những người đã đóng góp ý kiến, ý tưởng, và hỗ trợ trong dự án này. Sự đoàn kết và cống hiến của mọi người đã làm cho dự án trở thành một thành tựu đáng tự hào.

Dự án này đã mang lại cho chúng tôi rất nhiều kiến thức, kỹ năng, và trải nghiệm quý báu. Chúng em hy vọng rằng nó cũng sẽ có lợi ích và giá trị đối với mọi người như nó đã mang lại cho chúng em.

Xin chân thành cảm ơn mọi người một lần nữa.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

Chúng em xin cam đoan đây là sản phẩm đồ án của riêng chúng em và được sự hướng dẫn của thầy Lưu Giang Nam và thầy Trần Tấn Thành. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của chúng em. Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do chúng tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm 2024*

*Tác giả*

*(ký và ghi rõ họ tên)*

# PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

# TÓM TẮT

Mô hình VQ-VAE được triển khai trong Keras và TensorFlow để sinh ảnh, kết hợp giữa autoencoder và vector quantization để tạo ra một không gian biểu diễn rời rạc. Mô hình này thể hiện khả năng mô hình hóa dữ liệu rời rạc mạnh mẽ qua việc áp dụng trên bộ dữ liệu MNIST. Một điểm đặc biệt của VQ-VAE là không gian latent rời rạc, tạo điều kiện cho việc mô hình hóa dữ liệu có tính chất rời rạc như văn bản. Tuy nhiên, mô hình này đòi hỏi quá trình huấn luyện hai giai đoạn để học được prior phù hợp, là một thách thức so với việc lấy mẫu dễ dàng từ VAEs. Để giải quyết vấn đề sinh mẫu, một mô hình PixelCNN được huấn luyện như một prior phức tạp, cho phép sinh ảnh mới với chất lượng cao hơn.

# MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc167173885)

[PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN 3](#_Toc167173886)

[TÓM TẮT 4](#_Toc167173887)

[MỤC LỤC 5](#_Toc167173888)

[DANH MỤC KÝ HIỆU VÀ CHỮ VIẾT TẮT 6](#_Toc167173889)

[DANH MỤC CÁC HÌNH VẼ 7](#_Toc167173890)

[DANH MỤC CÁC BẢNG 8](#_Toc167173891)

[1. Giới thiệu về bài toán 9](#_Toc167173892)

[2. Phân tích bài toán và phương pháp thực hiện 9](#_Toc167173893)

[2.1. Yêu cầu của bài toán: 9](#_Toc167173894)

[2.2. Kiến trúc mô hình sử dụng 10](#_Toc167173895)

[2.2a. Mô hình VQ-VAE 10](#_Toc167173896)

[2.2b. Mô hình PixelCNN 12](#_Toc167173897)

[2.3. VQ-VAE + PixelCNN 14](#_Toc167173898)

[2.4. Ưu điểm việc kết hợp VQ-VAE và PixelCNN 14](#_Toc167173899)

[3. Thực nghiệm 15](#_Toc167173900)

[3.1. Dữ liệu 15](#_Toc167173901)

[3.2. Xử lý và tăng cường dữ liệu 15](#_Toc167173902)

[3.3. Công nghệ sử dụng: 15](#_Toc167173903)

[3.4. Cách đánh giá 16](#_Toc167173904)

[4. Kết quả đạt được 16](#_Toc167173905)

[4.1. Cài đặt 16](#_Toc167173906)

[4.1a. VQ-VAE 16](#_Toc167173907)

[4.1b. PixelCNN 17](#_Toc167173908)

[4.2. Kết quả đạt được 17](#_Toc167173909)

[4.2a. Mô hình VQ-VAE 17](#_Toc167173910)

[4.2b. Mô hình PixelCNN 18](#_Toc167173911)

[5. Tổng kết 19](#_Toc167173912)

[5.1. Kết luận 19](#_Toc167173913)

[5.2. Hạn chế của các phương pháp 20](#_Toc167173914)

[5.3. Hướng phát triển 20](#_Toc167173915)

[LÀM VIỆC NHÓM 22](#_Toc167173916)

[TÀI LIỆU THAM KHẢO 23](#_Toc167173917)

# 

# DANH MỤC KÝ HIỆU VÀ CHỮ VIẾT TẮT

**CÁC KÝ HIỆU:**

**CÁC CHỮ VIẾT TẮT**

# DANH MỤC CÁC HÌNH VẼ

[Hình 1: Bộ dữ liệu MNIST 10](#_Toc167173715)

[Hình 2: Mô hình VQ-VAE 11](#_Toc167173716)

[Hình 3: Mô hình PixelCNN 13](#_Toc167173717)

[Hình 4: Biểu đồ biểu diễn Losses và Codes norms 18](#_Toc167173718)

[Hình 5: Biểu đồ biểu diễn Loss và Accuracy 18](#_Toc167173719)

[Hình 6: Kết quả dự đoán khi kết hợp 2 mô hình 19](#_Toc167173720)

# DANH MỤC CÁC BẢNG

[Bảng 1: Công nghệ sử dụng cho thực nghiệm bài toán image generation 15](#_Toc167173863)

[Bảng 2: Các tham số của mô hình VQ-VAE 17](#_Toc167173864)

[Bảng 3: Các tham số của mô hình PixelCNN 17](#_Toc167173865)

**Keras VQ-VAE for image generation**

# 1. Giới thiệu về bài toán

Trong nghiên cứu này, chúng em tập trung vào việc khám phá khả năng sinh ảnh mới từ một không gian ẩn (latent space), thông qua việc áp dụng mô hình VQ-VAE, một biến thể đặc biệt của autoencoder. Mô hình này kết hợp phép lượng tử vector (vector quantization) để tạo ra một không gian biểu diễn rời rạc, từ đó cho phép chúng em sinh ra ảnh mới mang tính chất tương tự với dữ liệu huấn luyện. Điểm nhấn của nghiên cứu là việc sử dụng bộ giải mã PixelCNN, một mô hình sinh có khả năng mô hình hóa các phụ thuộc giữa các pixel dựa trên bối cảnh đã được sinh ra trước đó, thông qua các phép tích chập được kiểm soát và che giấu. Kết hợp hai mô hình này, chúng em mở ra một hướng tiếp cận mới trong việc sinh ảnh, không chỉ dựa vào việc tái tạo chính xác mà còn trên khả năng hiểu và tái hiện các mối quan hệ phức tạp giữa các pixel, từ đó tạo ra các ảnh mới với độ chính xác và tính thực tế cao. Điều này không chỉ góp phần vào lĩnh vực học sâu và thị giác máy tính mà còn mở ra các ứng dụng tiềm năng trong việc tạo ra dữ liệu huấn luyện cho các mô hình học máy, cải thiện hiệu suất trong các tác vụ nhận dạng và phân loại ảnh.

Ý nghĩa của bài toán:

Trong bài báo cáo này, chúng em sử dụng dataset mnist: [MNIST](https://www.kaggle.com/datasets/vikramtiwari/mnist-numpy/data), áp dụng mô hình Keras VQ-VAE, một phương pháp tiên tiến trong lĩnh vực học sâu và thị giác máy tính, để tạo ra ảnh mới từ không gian ẩn. Mô hình này, kết hợp với kỹ thuật Vector Quantization (VQ), cho phép sinh ra ảnh mới với đặc điểm tương tự như dữ liệu huấn luyện, mở ra khả năng ứng dụng mô hình trong nhiều lĩnh vực thực tiễn.

# 2. Phân tích bài toán và phương pháp thực hiện

## 2.1. Yêu cầu của bài toán:

Với bộ dữ liệu [MNIST](https://www.kaggle.com/datasets/vikramtiwari/mnist-numpy/data), bao gồm một tập hợp các ảnh kích thước 28x28 pixel, mỗi ảnh đại diện cho một chữ số từ 0 đến 9. Tập dữ liệu gốc được thu thập từ các mẫu chữ số viết tay của nhân viên và sinh viên tại Viện Tiêu chuẩn và Công nghệ Quốc gia Hoa Kỳ. Các ảnh đã được chuẩn hóa và chuyển đổi sang dạng xám (grayscale) để đơn giản hóa quá trình xử lý. MNIST chứa 60,000 ảnh trong tập huấn luyện và 10,000 ảnh trong tập kiểm tra. Dữ liệu đã được chia thành các tập này để đánh giá hiệu suất của mô hình trên dữ liệu không nhìn thấy trước.

A screenshot of a black board with white numbers

Description automatically generated

Hình 1: Bộ dữ liệu MNIST

Mục tiêu cụ thể của bài toán bao gồm:

Nhận dạng chính xác: Mô hình cần có khả năng nhận dạng chính xác các chữ số từ 0 đến 9 từ ảnh viết tay. Độ chính xác phân loại là chỉ số đánh giá chính.

Tự động hóa: Quá trình nhận dạng phải hoàn toàn tự động, không yêu cầu sự can thiệp của con người trong quá trình phân loại sau khi mô hình đã được huấn luyện.

Hiệu suất thời gian thực: Mô hình nên có khả năng đưa ra dự đoán nhanh chóng, phù hợp với các ứng dụng thời gian thực.

Khả năng tổng quát hóa: Mô hình phải hiệu quả không chỉ trên tập huấn luyện mà còn trên tập kiểm tra, đảm bảo rằng nó có khả năng tổng quát hóa tốt trên dữ liệu không nhìn thấy trước.

Robustness: Mô hình cần phải đủ mạnh để có thể xử lý các biến thể nhỏ trong cách viết, như kiểu chữ khác nhau hoặc nét viết mờ.

## 2.2. Kiến trúc mô hình sử dụng

### 2.2a. Mô hình VQ-VAE

Mô hình VQ-VAE (Vector Quantized-Variational AutoEncoder) là một biến thể của mô hình AutoEncoder truyền thống, được thiết kế để tạo ra một không gian biểu diễn rời rạc cho dữ liệu đầu vào thông qua kỹ thuật Vector Quantization (VQ). Điều này cho phép mô hình sinh ra dữ liệu mới với chất lượng cao và tính đa dạng từ không gian biểu diễn rời rạc này.

Mô hình VQ-VAE bao gồm ba thành phần chính: Encoder, Vector Quantizer, và Decoder.

A diagram of a bird

Description automatically generated

Hình 2: Mô hình VQ-VAE

1. Encoder:

Có hai cấp độ của bộ mã hóa (Encoder), "Top Level" và "Bottom Level", mỗi cấp độ thực hiện chức năng mã hóa dữ liệu đầu vào thành biểu diễn ẩn. Điều này cho thấy mô hình có thể sử dụng kiến trúc nhiều tầng hoặc nhiều mô-đun để xử lý thông tin ở các mức độ trừu tượng khác nhau.

1. Vector Quantizer (VQ):

Đây là bước trung gian giữa Encoder và Decoder, nơi biểu diễn ẩn được lượng tử hóa thành không gian rời rạc. Vector Quantizer chọn các vector từ một bộ codebook đã được học trước để đại diện cho biểu diễn ẩn, giúp cho việc tái tạo dữ liệu có thể được thực hiện một cách hiệu quả hơn.

1. Decoder:

Decoder nhận biểu diễn rời rạc từ VQ và tái tạo lại hình ảnh gốc. Trong hình, hai hình ảnh được hiển thị: "Original" là hình ảnh ban đầu, và "Reconstruction" là hình ảnh được tái tạo bởi Decoder sau khi qua quá trình mã hóa và lượng tử hóa.

### 2.2b. Mô hình PixelCNN

Mô hình PixelCNN (Pixel Convolutional Neural Network) là một kiến trúc mạng nơ-ron sinh tự hồi quy, được thiết kế để mô hình hóa phân phối xác suất của các ảnh pixel bằng pixel. Mô hình này sử dụng các lớp tích chập mặt nạ để đảm bảo rằng dự đoán của mỗi pixel chỉ dựa trên các pixel đã biết trước đó theo một thứ tự xác định, từ trên xuống dưới và từ trái sang phải. Điều này giúp PixelCNN có khả năng sinh ra ảnh mới một cách tuần tự và có điều kiện.

Các thành phần chính của mô hình PixelCNN bao gồm:

1. Lớp Tích Chập Mặt Nạ:

PixelCNN sử dụng các lớp tích chập mà các kernel (bộ lọc) của chúng được áp dụng mặt nạ để đảm bảo tính tự hồi quy. Mặt nạ này được thiết kế để kernel chỉ có thể truy cập vào các pixel ở vị trí trước đó trong thứ tự xác định, ngăn chặn thông tin từ các pixel chưa được xem xét "rò rỉ" vào quá trình dự đoán.

Có hai loại mặt nạ trong PixelCNN: Mặt nạ loại A và loại B. Mặt nạ loại A được sử dụng cho lớp đầu tiên để đảm bảo rằng không có thông tin nào từ pixel hiện tại được sử dụng trong dự đoán của chính nó, trong khi mặt nạ loại B cho phép pixel hiện tại được dự đoán dựa trên chính nó và các pixel trước đó.

1. Kiến trúc Tích Chập:

Mạng sử dụng các lớp tích chập chồng chéo nhau để tăng độ sâu và khả năng biểu diễn của mô hình. Các lớp tích chập này có thể bao gồm các lớp tích chập thông thường và lớp tích chập mặt nạ, được sắp xếp theo một cách thức nhất định để tối ưu hóa hiệu suất của mô hình trong việc mô hình hóa các phụ thuộc giữa các pixel.

1. Xử lý Đầu Ra:

Đầu ra của mạng thường được xử lý qua một hoặc nhiều lớp kết nối đầy đủ hoặc các lớp tích chập để sinh ra phân phối xác suất của các giá trị pixel tiếp theo. Mỗi pixel có thể được mô hình hóa bằng một phân phối đa thức để dự đoán xác suất của mỗi giá trị pixel có thể có.

A grid of numbers in different colors

Description automatically generated

Hình 3: Mô hình PixelCNN

* Phần Trái:

Biểu đồ Histogram: Phần này thể hiện phân bố cường độ của các pixel trong một hình ảnh. Trong PixelCNN, biểu đồ này có thể giúp hiểu được các giá trị pixel phổ biến và cách chúng được mô hình hóa trong quá trình dự đoán.

Mô hình Hóa: Hình minh họa cho thấy cách PixelCNN sử dụng thông tin từ các pixel đã biết (được chỉ ra bởi các đường màu đỏ) để dự đoán giá trị của pixel tiếp theo. Mô hình chỉ có thể điều kiện hóa trên các pixel đã được sinh ra trước đó, đảm bảo tính tự hồi quy của quá trình sinh ảnh.

* Phần Giữa:

Ma trận Mặt Nạ: Đây là ma trận được sử dụng để áp dụng mặt nạ lên bộ lọc 5x5 trong PixelCNN. Mục đích của việc này là để đảm bảo rằng mô hình không thể "nhìn thấy" các pixel nằm dưới hoặc ngay bên phải của pixel hiện tại mà nó đang cố gắng dự đoán. Điều này giúp ngăn chặn thông tin từ tương lai ảnh hưởng đến dự đoán hiện tại.

* Phần Phải:

Điểm Mù: Phần này chỉ ra rằng PixelCNN có một điểm mù trong trường nhận thức, nơi mà nó không thể sử dụng để làm ra các dự đoán. Điểm mù này là hậu quả của việc sử dụng mặt nạ trong các bộ lọc, hạn chế thông tin mà mô hình có thể truy cập khi dự đoán một pixel mới.

Chồng Các Lớp Tích Chập: Hình ảnh cuối cùng minh họa cách các lớp tích chập được chồng lên nhau (màu xanh và tím) để thu thập toàn bộ trường nhận thức. Việc sử dụng nhiều lớp tích chập giúp mô hình có thể hiểu và tích hợp thông tin từ một khu vực lớn hơn, cho phép nó dự đoán chính xác hơn dựa trên một lượng lớn các pixel liên quan.

## 2.3. VQ-VAE + PixelCNN

Kết hợp VQ-VAE (Vector Quantized-Variational AutoEncoder) và PixelCNN tạo nên một hệ thống mạnh mẽ trong việc sinh ảnh, cho phép tạo ra các biểu diễn ảnh mới từ không gian ẩn rời rạc mà VQ-VAE cung cấp. Hệ thống này tận dụng lợi thế của cả hai mô hình để tạo ra các biểu diễn ngữ nghĩa đa dạng và chi tiết ảnh pixel-by-pixel chất lượng cao. Quá trình hoạt động của VQ-VAE + PixelCNN:

1. VQ-VAE:

Encoder: Chuyển đổi ảnh gốc thành một biểu diễn ẩn liên tục.

Vector Quantizer: Lượng tử hóa biểu diễn ẩn này thành các vector mã rời rạc, mỗi vector chứa thông tin quan trọng về các đặc trưng của ảnh gốc.

Decoder: Sử dụng các vector mã để tái tạo lại ảnh gốc, giúp mô hình học cách giảm thiểu sai số giữa ảnh gốc và ảnh tái tạo.

1. PixelCNN:

Mô hình sinh: Sử dụng các vector mã từ VQ-VAE, PixelCNN đóng vai trò là mô hình sinh để dự đoán màu sắc của từng điểm ảnh mới, dựa trên các điểm ảnh đã được sinh ra trước đó.

Tính tự hồi quy: PixelCNN hoạt động theo cách tuần tự, dự đoán từng pixel một, cho phép tạo ra các chi tiết ảnh một cách chính xác và tuần tự.

## 2.4. Ưu điểm việc kết hợp VQ-VAE và PixelCNN

Kết hợp VQ-VAE và PixelCNN tạo nên một hệ thống mạnh mẽ, tận dụng lợi thế của cả hai mô hình để tạo ra một giải pháp sinh ảnh và mô hình hóa không gian ẩn với hiệu quả cao. Dưới đây là các ưu điểm chính khi kết hợp hai mô hình này:

1. Tạo ra không gian ẩn rời rạc và chính xác:

VQ-VAE tạo ra một không gian ẩn rời rạc, giúp đơn giản hóa việc mô hình hóa phân phối của dữ liệu. Không gian ẩn rời rạc này cho phép PixelCNN dễ dàng học và dự đoán các vector mã hóa tiếp theo, từ đó tăng cường độ chính xác và tính ổn định của mô hình sinh.

1. Cải thiện chất lượng sinh ảnh:

PixelCNN, khi được sử dụng như một mô hình sinh, có thể tạo ra các ảnh mới từ các vector mã được học bởi VQ-VAE. Sự kết hợp này cho phép sinh ra ảnh không chỉ với độ chính xác cao mà còn với chi tiết rõ ràng và sắc nét, nhờ vào khả năng của PixelCNN trong việc mô hình hóa các điều kiện phụ thuộc giữa các pixel.

1. Tận dụng lợi thế của mô hình tự hồi quy:

PixelCNN là một mô hình tự hồi quy, cho phép nó tận dụng thông tin từ các pixel trước đó để dự đoán pixel tiếp theo. Điều này, khi kết hợp với các vector mã từ VQ-VAE, giúp tạo ra các biểu diễn ảnh mới một cách tuần tự và chính xác, làm tăng khả năng sinh ảnh và tái tạo ảnh từ không gian ẩn.

1. Giảm thiểu mất mát trong quá trình lượng tử hóa:

Trong khi VQ-VAE có thể gây ra một số mất mát thông tin do quá trình lượng tử hóa, việc sử dụng PixelCNN giúp khắc phục phần nào vấn đề này bằng cách học cách sinh dữ liệu từ không gian rời rạc một cách hiệu quả, từ đó giảm thiểu tác động của mất mát thông tin và cải thiện chất lượng của ảnh tái tạo.

1. Khả năng ứng dụng rộng rãi:

Sự kết hợp của VQ-VAE và PixelCNN không chỉ hữu ích cho việc sinh ảnh mà còn mở rộng được ứng dụng trong nhiều lĩnh vực khác như sinh âm thanh, xử lý ngôn ngữ tự nhiên, và các bài toán phức tạp khác trong học sâu, nơi mà việc mô hình hóa các phân phối phức tạp là cần thiết.

# 3. Thực nghiệm

## 3.1. Dữ liệu

Đồ án tiến hành thực nghiệm trên bộ dữ liệu MNIST. Tập dữ liệu MNIST bao gồm một tập hợp các ảnh kích thước 28x28 pixel, mỗi ảnh đại diện cho một chữ số từ 0 đến 9. Tập dữ liệu gốc được thu thập từ các mẫu chữ số viết tay của nhân viên và sinh viên tại Viện Tiêu chuẩn và Công nghệ Quốc gia Hoa Kỳ. Các ảnh đã được chuyển đổi sang dạng xám (grayscale) và chúng ta chuẩn hóa các giá trị màu để các đầu vào nằm trong khoảng [0., 1.] để đơn giản hóa quá trình xử lý.

## 3.2. Xử lý và tăng cường dữ liệu

## 3.3. Công nghệ sử dụng:

|  |  |
| --- | --- |
| **Ngôn ngữ** | Python |
| **Thư viện** | Tensorflow,Keras,matplotlib,numpy |
| **Môi trường** | Kaggle |

Bảng 1: Công nghệ sử dụng cho thực nghiệm bài toán image generation

## 3.4. Cách đánh giá

Về đánh giá chúng tôi chọn đánh giá dựa trên 3 hàm loss là latent loss, mse loss (Mean Squared Error), SparseCategoricalCrossentropy loss và accuracy:

* Latent Loss: Mô hình VQ-VAE cố gắng biểu diễn dữ liệu đầu vào bằng cách mã hóa nó thành một không gian ẩn (latent space) và sau đó lượng tử hóa vector từ không gian ẩn này. Mục tiêu của latent loss là đảm bảo rằng quá trình lượng tử hóa được thực hiện chính xác và hiệu quả. Latent loss giúp đảm bảo rằng quá trình lượng tử hóa được thực hiện chính xác và hiệu quả, tức là mỗi vector đầu vào được ánh xạ vào một vector gần nhất trong không gian mã lượng tử hóa.
* SparseCategoricalCrossentropy loss: Đo lường sự khác biệt giữa phân phối xác suất dự đoán và phân phối xác suất thực tế của các lớp. Hàm loss này phù hợp với các tình huống khi nhãn được biểu diễn dưới dạng các số nguyên (sparse labels) thay vì one-hot vectors. Cross-entropy là một phép đo thông thường và hiệu quả cho bài toán phân loại, đặc biệt khi kết hợp với hàm softmax để tính toán xác suất.
* MSE Loss: Là một phép đo thông thường để đánh giá mức độ sai lệch giữa các giá trị dự đoán và các giá trị thực tế trong các bài toán hồi quy. MSE tính toán tổng của bình phương của sự sai lệch giữa giá trị dự đoán và giá trị thực tế, sau đó chia cho số lượng mẫu để lấy giá trị trung bình.Công thức tính MSE như sau:

MSE =

Trong đó:

* + - n là số lượng mẫu
    - là giá trị thực tế của mẫu thứ i
    - là giá trị dự đoán của mẫu thứ i
* Accuracy: Khi xây dựng mô hình phân loại chúng ta sẽ muốn biết một cách khái quát tỷ lệ các trường hợp được dự báo đúng trên tổng số các trường hợp là bao nhiêu. Tỷ lệ đó được gọi là độ chính xác. Độ chính xác giúp ta đánh giá hiệu quả dự báo của mô hình trên một bộ dữ liệu. Độ chính xác càng cao thì mô hình của chúng ta càng chuẩn xác. Độ chính xác được tính theo công thức:

# 4. Kết quả đạt được

## 4.1. Cài đặt

### 4.1a. VQ-VAE

Mô hình VQ-VAE chỉnh sửa được huấn luyện với các thông số như sau:

|  |  |
| --- | --- |
| Learning rate | 3e-4 |
| Latent loss | True |
| MSE loss | True |
| Layer | [16, 32] |
| Optimizer | Adam |
| Epochs | 20 |
| Batch size | 128 |

Bảng 2: Các tham số của mô hình VQ-VAE

### 4.1b. PixelCNN

Mô hình PixelCNN chỉnh sửa được huấn luyện với các thông số như sau:

|  |  |
| --- | --- |
| Learning rate | 3e-4 |
| Epochs | 10 |
| SparseCategoricalCrossentropy Loss | True |
| Optimizer | Adam |
| Batch size | 128 |
| Blocks | 12 |
| Feature Maps | 32 |

Bảng 3: Các tham số của mô hình PixelCNN

## 4.2. Kết quả đạt được

### 4.2a. Mô hình VQ-VAE

Việc cung cấp 2 hàm loss cho mô hình VQ-VAE giúp đảm bảo rằng mô hình có khả năng tái tạo dữ liệu gốc từ dữ liệu được mã hóa, đảm bảo quá trình mã hóa và lượng tử hóa được thực hiện một cách hiệu quả và đúng đắn. Các metric ‘zq\_norm’ và ‘ze\_norm’ giúp đánh giá hiệu suất của mô hình bằng cách đo lường kích thước của các vector được mã hóa và lượng tử hóa. Các metric này có thể giúp giám sát và đánh giá sự biến đổi của các vector trong quá trình huấn luyện và đảm bảo rằng mô hình đang học một cách hiệu quả.

A graph of loss and loss

Description automatically generated

Hình 4: Biểu đồ biểu diễn Losses và Codes norms

### 4.2b. Mô hình PixelCNN

Việc xây dựng mô hình PixelCNN cho bài toán image generation đã giúp mô hình VQ-VAE tạo ra các mã lượng tử hiệu quả và chất lượng hơn.

A graph of a line

Description automatically generated with medium confidence

Hình 5: Biểu đồ biểu diễn Loss và Accuracy

Tiến hành dự đoán một số mẫu. Các kết quả cho thấy sự kết hợp hoàn hảo của 2 mô hình VQ-VAE và PixelCNN khi tạo ra các ảnh mới có chất lượng tốt hơn từ không gian latent.

A number in squares with different colors

Description automatically generated with medium confidence

Hình 6: Kết quả dự đoán khi kết hợp 2 mô hình

# 5. Tổng kết

## 5.1. Kết luận

**VQ-VAE:** Mô hình được sử dụng để học một biểu diễn hiệu quả và phân tách các đặc trưng của ảnh. Nó mã hóa ảnh đầu vào thành các vector latent rời rạc, tạo ra một không gian biểu diễn dữ liệu hiệu quả hơn.

**PixelCNN:** Mô hình PixelCNN được sử dụng để học một prior phân phối xác suất trên không gian latent của VQ-VAE. Điều này cho phép tạo ra các mẫu mới từ không gian latent thông qua việc lấy mẫu từ phân phối xác suất đã học được.

Bằng cách kết hợp hai mô hình này, mục đích là tạo ra các ảnh mới có chất lượng tốt từ không gian latent của mô hình VQ-VAE. Mô hình PixelCNN học cách tạo ra các ảnh mới từ không gian latent này, cung cấp sự đa dạng và tính chất tự nhiên cho các ảnh được sinh ra

## 5.2. Hạn chế của các phương pháp

Mặc dù phương pháp kết hợp hai mô hình VQ-VAE và PixelCNN cho tập dữ liệu MNIST có thể đem lại nhiều lợi ích, nhưng cũng tồn tại một số hạn chế:

* Tính phức tạp của huấn luyện: Huấn luyện cả hai mô hình VQ-VAE và PixelCNN đều đòi hỏi một lượng lớn dữ liệu và thời gian tính toán đáng kể. Đặc biệt, mạng PixelCNN thường có thời gian huấn luyện dài do tính chậm của quá trình sampling pixel by pixel.
* Chất lượng ảnh sinh ra: Mặc dù mục tiêu là tạo ra các ảnh mới có chất lượng cao, nhưng trong một số trường hợp, các ảnh sinh ra vẫn có thể bị mờ hoặc không chính xác. Điều này có thể do sự phức tạp của bài toán sinh ảnh và khó khăn trong việc học được phân phối xác suất chính xác cho không gian latent.
* Khả năng tái tạo chính xác: Mặc dù mô hình có thể tạo ra các ảnh mới, nhưng không đảm bảo rằng các ảnh này sẽ chính xác đến từng chi tiết so với dữ liệu gốc. Có thể có sự mất mát thông tin trong quá trình mã hóa và tái tạo.
* Độ phức tạp của mô hình: Sự kết hợp của hai mô hình VQ-VAE và PixelCNN tạo ra một hệ thống phức tạp với nhiều tham số cần được điều chỉnh và tinh chỉnh. Điều này có thể làm tăng chi phí tính toán và đòi hỏi nhiều kỹ năng chuyên môn để triển khai và điều chỉnh.
* Khả năng tổng quát hóa: Hiệu suất của mô hình có thể giảm khi áp dụng cho dữ liệu ngoài tập dữ liệu huấn luyện, đặc biệt là khi mô hình gặp phải các dạng dữ liệu mới hoặc phức tạp hơn.

## 5.3. Hướng phát triển

Một số hướng phát triển tiềm năng cho bài toán kết hợp hai mô hình VQ-VAE và PixelCNN cho tập dữ liệu MNIST có thể bao gồm:

* Tăng cường tính đa dạng của ảnh sinh ra: Tìm cách tăng cường tính đa dạng và sự tự nhiên của các ảnh sinh ra bằng cách sử dụng các phương pháp như augmentations, regularization, hoặc đầu vào nhiễu để mô hình học được các biểu diễn phong phú hơn và tạo ra các ảnh mới đa dạng hơn.
* Tối ưu hiệu suất và tốc độ: Tìm cách tối ưu hiệu suất và tốc độ của quá trình huấn luyện và sinh ảnh. Điều này có thể bao gồm việc sử dụng kỹ thuật huấn luyện song song, tối ưu hóa mã hóa và giải mã, hoặc sử dụng các kiến trúc mạng hiệu quả hơn.
* Nâng cao chất lượng ảnh sinh ra: Nghiên cứu và phát triển các phương pháp mới để cải thiện chất lượng và sự chính xác của các ảnh sinh ra, bao gồm việc sử dụng các mô hình sinh ảnh phức tạp hơn, kiến trúc mạng PixelCNN cải tiến, hoặc các kỹ thuật khác để giảm mất mát thông tin trong quá trình tái tạo.
* Mở rộng áp dụng cho các tập dữ liệu khác: Nghiên cứu và phát triển phương pháp để mở rộng áp dụng của mô hình cho các tập dữ liệu khác ngoài MNIST, bao gồm các bài toán về ảnh có độ phức tạp cao hơn hoặc đa dạng hơn.
* Khảo sát về hiệu quả của prior học được: Nghiên cứu về cách mà prior được học từ PixelCNN ảnh hưởng đến chất lượng và đa dạng của các ảnh sinh ra, và tìm cách tối ưu hoá quá trình học prior để đạt được kết quả tốt nhất.
* Tích hợp các phương pháp khác nhau: Nghiên cứu và tích hợp các phương pháp mới và cải tiến để cải thiện hiệu suất và chất lượng của mô hình, bao gồm việc kết hợp với các mô hình sinh ảnh khác, kỹ thuật tái tạo ảnh, hoặc các kiến trúc mạng khác.

# LÀM VIỆC NHÓM

Bảng phân chia nhiệm vụ:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Công việc | Người thực hiện | Trạng thái công việc |
| 1 | Tìm, thu thập dữ liệu | Huy,Giang | 100% |
| 2 | Xây dựng mô hình | Huy,Giang | 100% |
| 3 | Huấn luyện mô hình | Huy,Giang | 100% |
| 4 | Viết báo cáo | Huy,Giang | 100% |

Cách thức làm việc nhóm: Offline kết hợp online.

Thời gian làm việc nhóm: Họp và làm việc mỗi tuần 1-3 buổi tại trọ và qua discord (1-2h mỗi lần)

# TÀI LIỆU THAM KHẢO

1.Sayak Paul.Vector-Quantized Variational Autoencoders. *keras.io/examples/generative/vq\_vae/.*

*2.*Aaron van den Oord et, al.Conditional Image Generation with PixelCNN Decoders. *arXiv:1606.05328*.

3.*Aaron van den Oord et, al.*Neural Discrete Representation Learning. *arXiv:1711.00937.*

4.Ali Razavi et, al. Generating Diverse High-Fidelity Images with VQ-VAE 2. *arXiv:1906.00446*.

5.Aayush Bansal et, al. PixelCNN: Example-based Image Synthesis. *arXiv:1708.05349*.

6.Akash Saravanan et, al. Pixel VQ-VAEs for improved Pixel Art Representation. *arXiv:2203.12130*.