Ứng dụng Gen AI trong Tạo Nội dung Đa phương tiện

Mô tả tổng quan về bài toán chuyển đổi văn bản thành hình ảnh, âm thanh, và video.

Họ và tên: Vũ Thị Minh Thư
MSV: 22028116

Giới thiệu sản phẩm

- Sản phẩm: Tạo Nội Dung Tin Tức Từ Văn Bản Gốc bằng GenAI
- Mục tiêu: Tạo ra một trải nghiệm đa phương tiện giúp người dùng dễ dàng tiếp nhận thông tin nhanh chóng và trực quan.
- Kết quả: Một video, một trang HTML hiển thị toàn bộ tin tức.
- Lợi ích:
    + Tăng cường khả năng tiếp cận thông tin qua nội dung ngắn gọn, dễ hiểu và trực quan.
    + Giúp tiết kiệm thời gian khi theo dõi tin tức với các yếu tố đa phương tiện sinh động.

Giới thiệu sản phẩm

Sản phẩm tự động chuyển đổi nội dung tin tức từ văn bản gốc thành các định dạng trực quan và sinh động như hình ảnh, âm thanh, và video.

Quy trình thực hiện

1. Tóm tắt văn bản: Từ văn bản gốc, tạo ra nội dung tóm tắt ngắn gọn.
2. Tạo hình ảnh minh họa từ văn bản tóm tắt.
3. Chuyển văn bản thành giọng nói để tạo ra âm thanh.
4. Kết hợp hình ảnh, văn bản cuộn và âm thanh để tạo video.
5. Tạo trang HTML hiển thị tin tức với hình ảnh và nội dung văn bản.

Summarization - BART

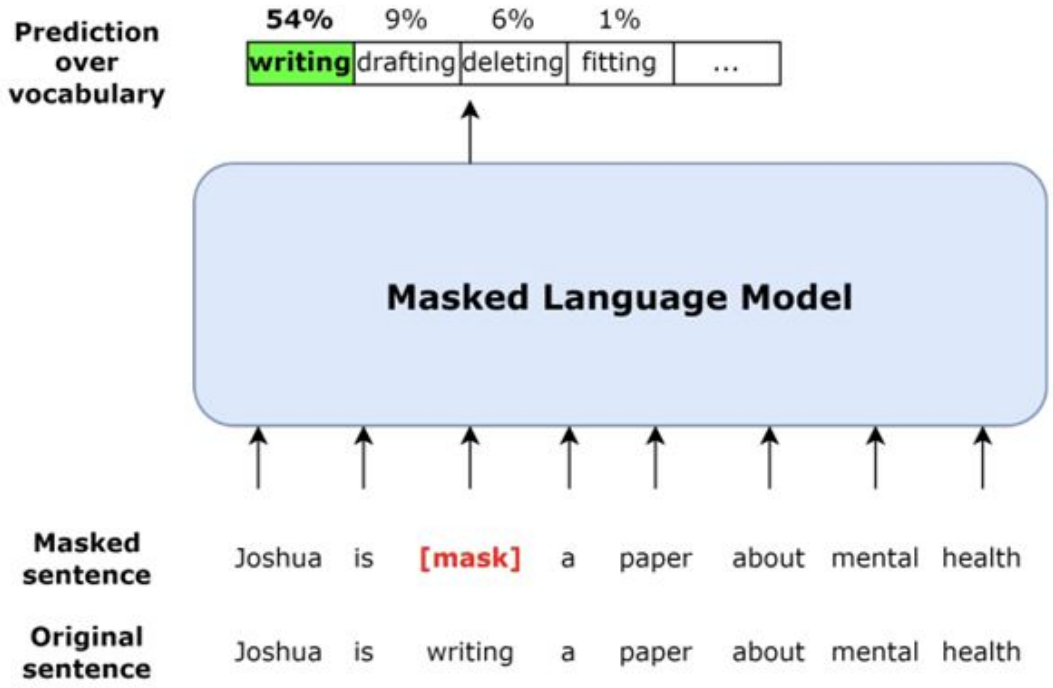
- **Mô hình BART (Bidirectional and Auto-Regressive Transformers)**: Được thiết kế cho các nhiệm vụ NLP, bao gồm tóm tắt văn bản.
- **Ứng dụng**: Giúp biến các đoạn văn bản dài thành những câu tóm tắt ngắn gọn, phù hợp để minh họa trong các video ngắn.
- **Hiệu quả của BART trong Nghiên cứu**:
    - Nghiên cứu của Facebook AI đã chứng minh rằng BART đạt hiệu quả cao khi tóm tắt văn bản trên bộ dữ liệu CNN/Daily Mail.
    - BART vượt trội hơn các mô hình trước đó nhờ khả năng duy trì ý nghĩa và ngữ cảnh khi tạo văn bản tóm tắt.
    - Kết quả nghiên cứu cho thấy BART là một trong những mô hình hàng đầu về tóm tắt văn bản tự động.

Summarization - BART

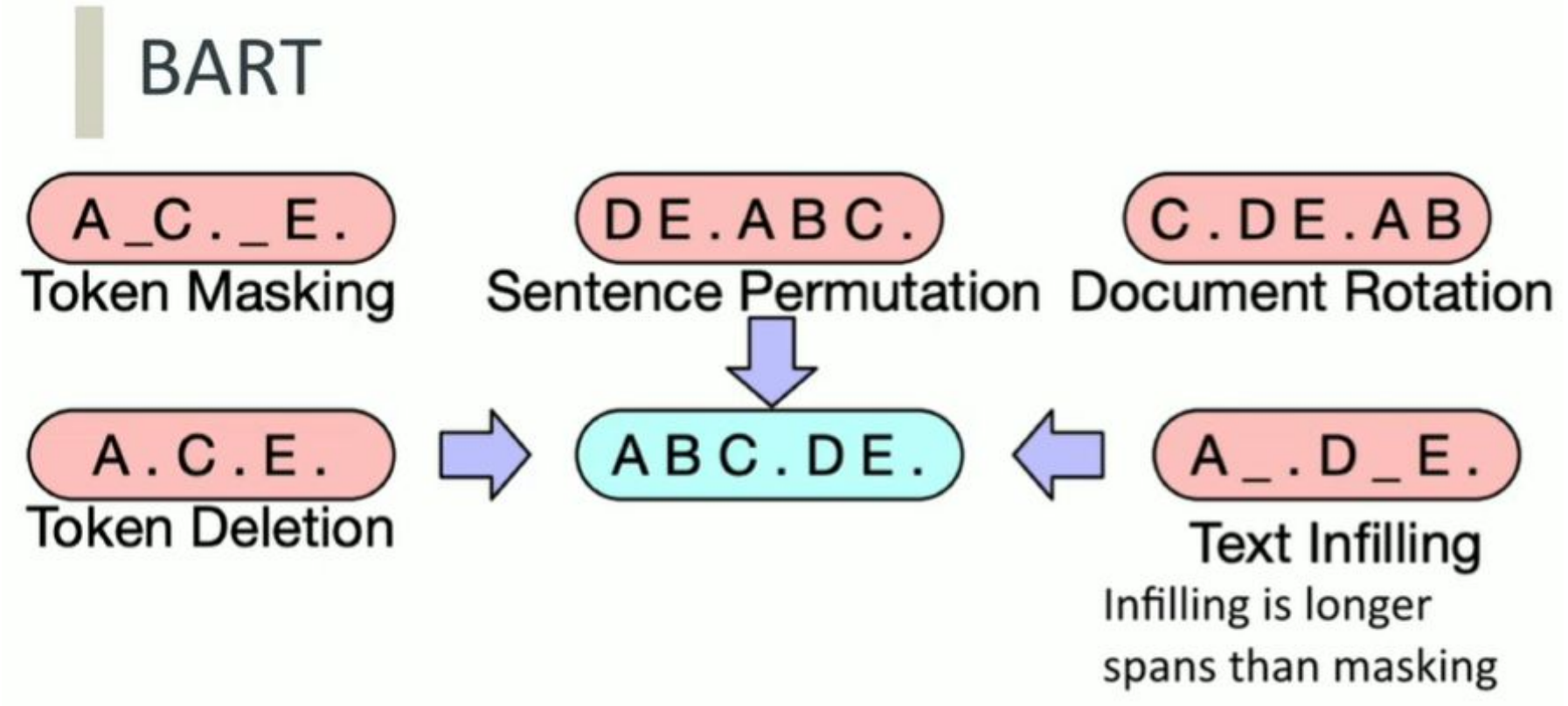
- **Masking**: Để mô hình học cách dự đoán các từ bị che để tái hiện lại nội dung ban đầu, giúp BART học được ngữ nghĩa và ngữ cảnh.

- **Pre-training**: Một số token (10–15%) trong văn bản sẽ bị che ngẫu nhiên. Mô hình được huấn luyện để dự đoán chính xác các token bị che này.

- Sản phẩm dùng **facebook/bart-large-cnn,** là phiên bản mở rộng của BART với khoảng 400 triệu tham số, được tối ưu hóa cho tác vụ tóm tắt văn bản



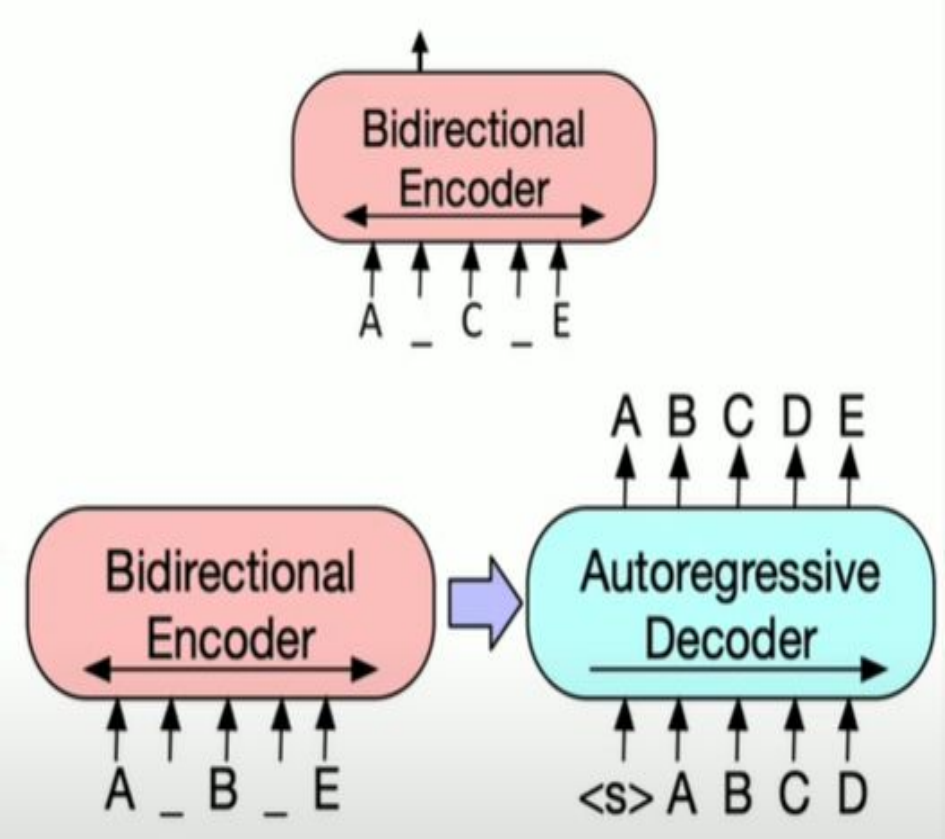
Cơ chế hoạt động của BART



Cơ chế hoạt động của BART

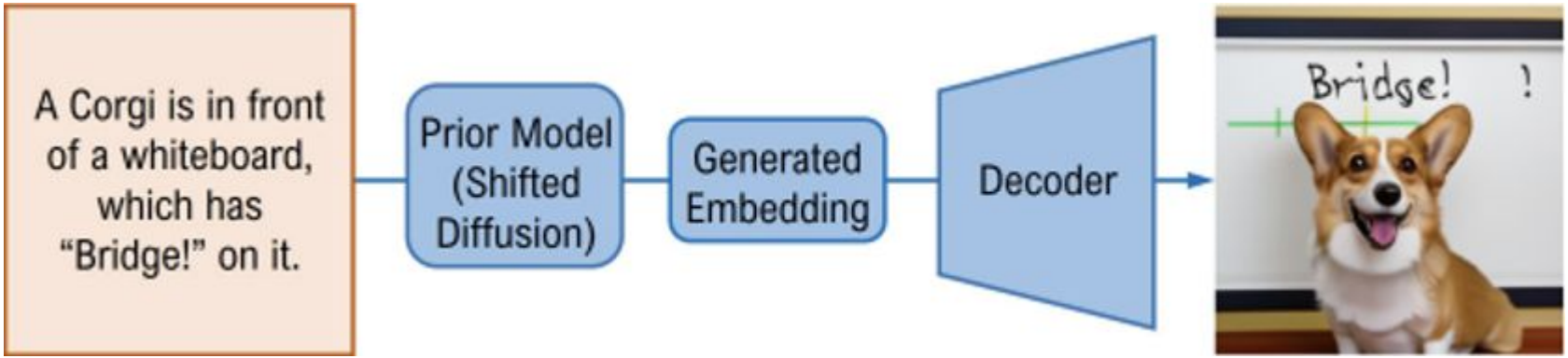
- Mô hình là một kiến trúc encoder-decoder kết hợp khả năng của cả mô hình masking như BERT và các cơ chế tạo chuỗi của GPT (Generative Pre-trained Transformer).

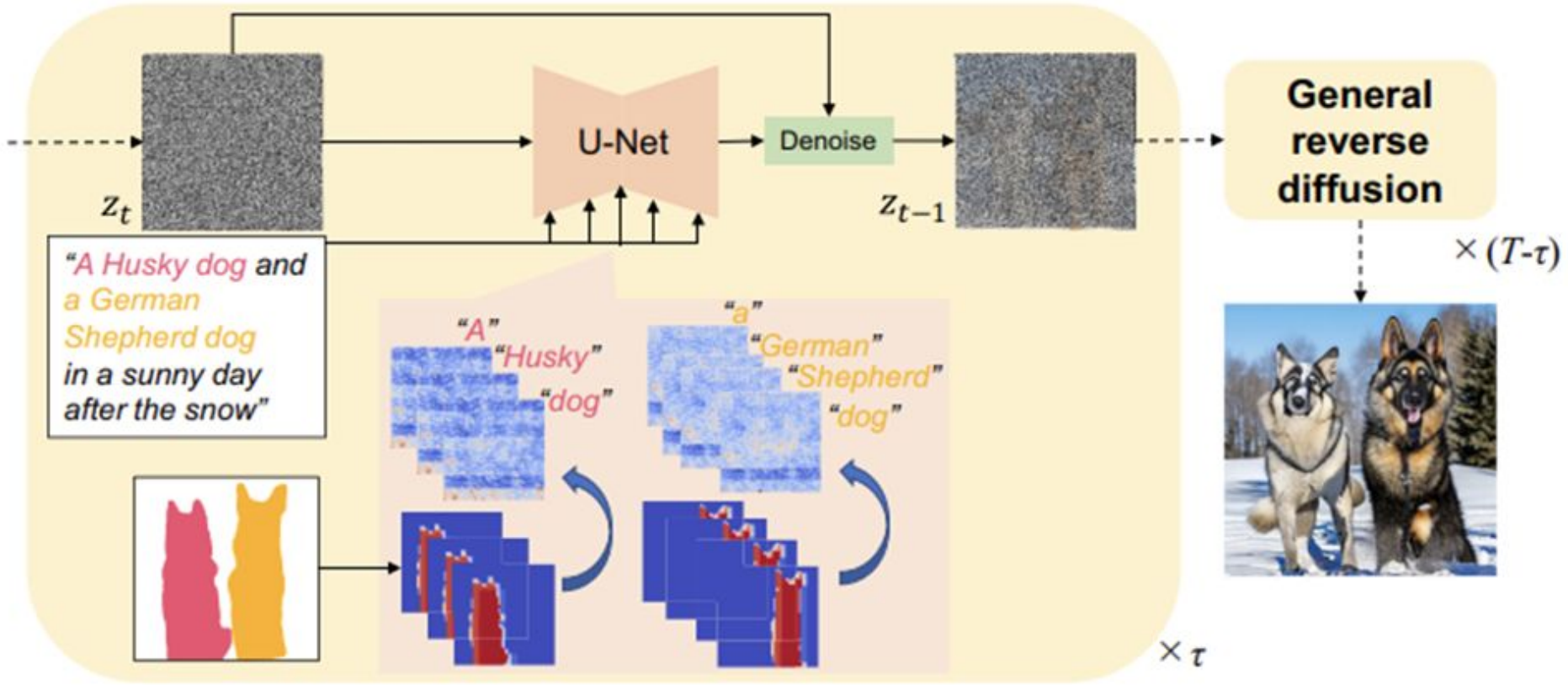
- Cấu trúc hai chiều và tự hồi quy (Bidirectional and Auto-Regressive) cho phép mô hình BART hiểu và tạo văn bản một cách liền mạch, phù hợp cho các tác vụ như tóm tắt văn bản, dịch máy, và tạo văn bản.



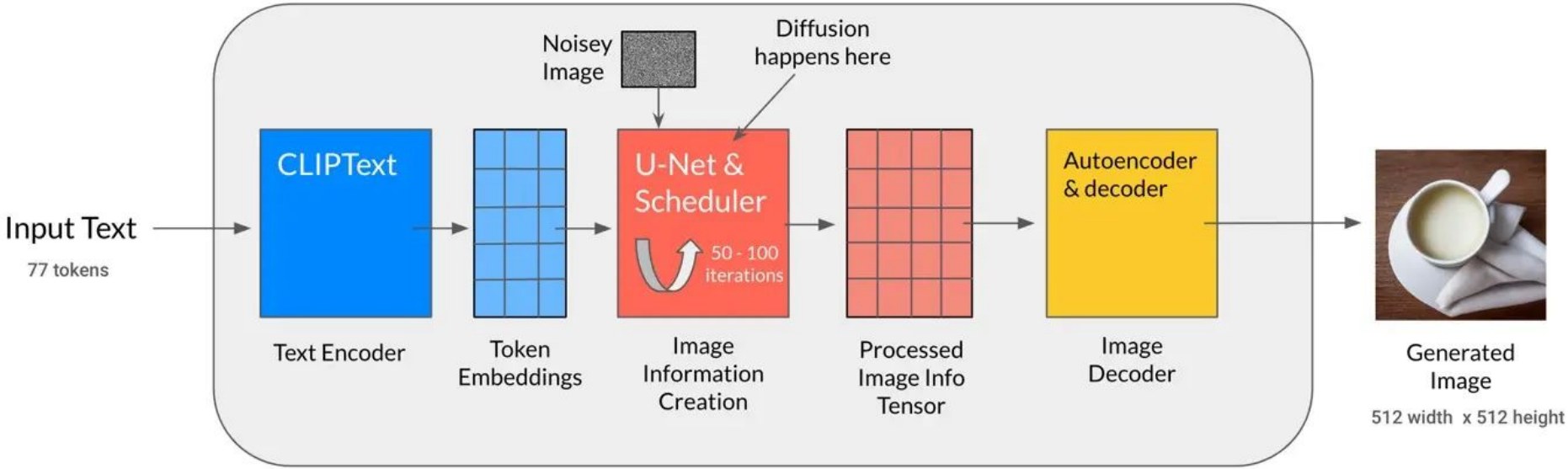
Text-to-Image Generation - Stable Diffusion Model

- **Mô hình Stable Diffusion**: Một mạng neural khuếch tán tiên tiến, học từ hàng trăm triệu hình ảnh kèm chú thích văn bản. Thông qua quá trình học, Stable Diffusion có thể hiểu và liên kết nội dung giữa ngôn ngữ và hình ảnh một cách hiệu quả.
- **Cơ chế**: Khuếch tán ngược, mô hình bắt đầu với một ảnh hoàn toàn nhiễu, sau đó từng bước loại bỏ nhiễu dựa trên nội dung mô tả từ chuỗi văn bản. Qua quá trình này, mức độ nhiễu giảm dần, cho phép mô hình tổng hợp hình ảnh từ các yếu tố ngữ nghĩa phức tạp và tạo ra chi tiết chính xác, dần dần hình thành một bức ảnh có cấu trúc và ý nghĩa rõ ràng.




z_t
U-Net
Denoise
z_{t-1}
General reverse diffusion
$\times (T\text{-}\tau)$
"A Husky dog and a German Shepherd dog in a sunny day after the snow"
"A"
"Husky"
"dog"
"a"
"German"
"Shepherd"
"dog"
$\times \tau$

Stable Diffusion Architecture



Google Text-to-Speech (gTTS)

Google TTS: Công cụ chuyển đổi văn bản thành giọng nói của Google, sử dụng để tạo giọng nói tự nhiên, đáp ứng nhiều ngữ cảnh khác nhau. Cung cấp trải nghiệm tương tác tự nhiên, giảm thiểu cảm giác giọng đọc nhân tạo.

Tính năng nổi bật:

- Hỗ trợ nhiều ngôn ngữ và phong cách phát âm.
- Độ chính xác cao, giúp chuyển văn bản thành giọng nói một cách tự nhiên và mượt mà.

Công nghệ học sâu: Sử dụng các mô hình học sâu tiên tiến để điều chỉnh ngữ điệu và âm sắc, mang lại cảm giác tự nhiên và gần gũi cho người nghe

Minh chứng: Trợ lý ảo Google Assistant(dùng gTTS) chuyển đổi câu lệnh thành giọng nói mượt mà, hỗ trợ nhiều ngôn ngữ và phong cách phát âm.

Cơ chế Hoạt động của Google TTS

Chuyển đổi ký tự thành ngữ âm:

- Phân tích văn bản, xác định cấu trúc câu và từ.
- Áp dụng các quy tắc ngữ âm để chuyển đổi ký tự thành âm vị, chuẩn bị cho quá trình tạo âm thanh.

Điều chỉnh ngữ điệu và ngữ cảnh:

- Sau khi chuyển đổi thành âm vị, hệ thống điều chỉnh ngữ điệu và tốc độ phù hợp với ngữ cảnh.
- Kết quả là giọng nói tự nhiên, dễ hiểu và mượt mà

Create Video with Scrolling Text and Audio-MoviePy

- **Ghép nối Hình ảnh, Văn bản và Âm thanh**: MoviePy kết hợp từng thành phần (ảnh, văn bản, âm thanh) để tạo clip ngắn. Các clip này sau đó được ghép nối thành video hoàn chỉnh, tạo trải nghiệm xem đồng nhất.

- **Cuộn Văn bản**: Hiệu ứng cuộn văn bản di chuyển từ phải sang trái giúp văn bản nổi bật như trong các video tin tức chuyên nghiệp.

Kết quả và Tính năng của Sản phẩm

- **Video tin tức hoàn chỉnh**: Sản phẩm tự động tạo video với văn bản, hình ảnh và âm thanh, mang đến nội dung sinh động và lôi cuốn.
- **Trang HTML**: Hiển thị tin tức dưới dạng từng mục với hình ảnh và văn bản kèm theo
- **Trải nghiệm hấp dẫn với GenAI**: Sản phẩm mang đến trải nghiệm trực quan, dễ tiếp cận thông qua công nghệ AI thế hệ mới.

Tính mới và tính sáng tạo

Tự động hóa toàn diện quy trình từ văn bản đến đa phương tiện:

- **Điểm mới**: Hệ thống tích hợp chặt chẽ các công cụ GenAI để tự động chuyển đổi văn bản tin tức thành các định dạng đa phương tiện dễ tiếp nhận như hình ảnh, âm thanh và video.

- **Điểm sáng tạo**: Thay vì chỉ cung cấp văn bản hoặc hình ảnh đơn lẻ, hệ thống tạo ra một quy trình liền mạch chuyển đổi nội dung văn bản thành trải nghiệm nghe, nhìn sống động. Người dùng có thể tiếp cận tin tức qua nhiều định dạng mà không cần đọc toàn bộ văn bản gốc.

Tính mới và tính sáng tạo

Ứng dụng đồng thời nhiều mô hình AI tiên tiến:

- **Tính mới**: Kết hợp các mô hình AI tiên tiến như BART cho tóm tắt văn bản, Stable Diffusion để tạo hình ảnh từ văn bản, và Google TTS cho chuyển đổi văn bản thành giọng nói. Đây là sự tích hợp đồng bộ giữa NLP, AI hình ảnh và AI giọng nói trong một sản phẩm duy nhất.

- **Tính sáng tạo**: Mỗi mô hình AI đảm nhận một bước cụ thể, giúp tối ưu hóa chất lượng và tính tự nhiên của từng thành phần trong quy trình. Ví dụ, BART rút gọn văn bản nhưng vẫn giữ được ý nghĩa, Stable Diffusion tạo ra hình ảnh minh họa phong phú, và Google TTS mang lại giọng đọc tự nhiên cho nội dung. Sự kết hợp này đảm bảo tính chính xác và tính trực quan cao cho sản phẩm.

Tính mới và tính sáng tạo

Hiệu ứng cuộn văn bản và ghép nối đa phương tiện:

- **Điểm mới**: Hệ thống không chỉ kết hợp hình ảnh và âm thanh mà còn sử dụng hiệu ứng cuộn văn bản trong video, tạo trải nghiệm xem tin tức sống động như trên các kênh truyền thông chuyên nghiệp.

- **Điểm sáng tạo**: Hiệu ứng cuộn văn bản được đồng bộ với giọng đọc, giúp người xem dễ dàng theo dõi nội dung và tiếp nhận thông tin âm thanh một cách liền mạch. Đây là điểm nổi bật khiến sản phẩm trở nên thu hút hơn so với các hình thức trình bày tin tức truyền thống.

Tính mới và tính sáng tạo

Khả năng mở rộng và ứng dụng linh hoạt:

- **Tính mới**: Cấu trúc của giải pháp cho phép dễ dàng mở rộng để ứng dụng trong nhiều lĩnh vực khác nhau như học tập, báo cáo doanh nghiệp và quảng cáo.

- **Tính sáng tạo**: Giải pháp có thể được tùy chỉnh phù hợp với nhiều ngữ cảnh. Ví dụ, trong giáo dục, giải pháp có thể tự động tạo video bài giảng từ tài liệu văn bản; trong doanh nghiệp, có thể tạo các báo cáo kinh doanh đa phương tiện. Điều này mang lại các ứng dụng rộng rãi và linh hoạt cho nhiều ngành.

Kết luận và Tiềm năng Phát triển

Sản phẩm ứng dụng các mô hình GenAI tiên tiến để tự động hóa quy trình tạo nội dung tin tức đa phương tiện, mang lại trải nghiệm thông tin hiện đại và hấp dẫn.

Tiềm năng phát triển:

- Tăng cường tính cá nhân hóa nội dung, phù hợp với nhu cầu từng người dùng.
- Phát triển thêm các tính năng hỗ trợ ngôn ngữ địa phương.
- Cải thiện tốc độ xử lý và độ chính xác trong tóm tắt văn bản.

Lợi ích: Giải pháp giúp tiết kiệm thời gian, nâng cao trải nghiệm người dùng và cung cấp một cách thức tiếp cận tin tức mới mẻ, thú vị.

Demo

Link demo: https://youtu.be/6tBNfj3o1r4

Link báo cáo sản phẩm: https://youtu.be/Ml38-QkOmtw