THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/6sDWk28-4k0*
- Link slides (dạng .pdf đặt trên Github): *https://github.com/hoahk-uitsdh17/CS2205.APR2023/blob/main/Kh%C3%A1nh%20H%C3%B2a%20Hu%E1%BB%B3nh%20-%20xCS2205.DeCuong.FinalReport.Template.Slide.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Huỳnh Khánh Hòa
• MSSV: 220101005
 
	• Lớp: CS2205.APR2023
	• Tự đánh giá (điểm tổng kết môn): 9/10
	• Số buổi vắng: 0
	• Số câu hỏi QT cá nhân:
	• Số câu hỏi QT của cả nhóm:
	• Link Github: https://github.com/hoahk-uitsdh17/CS2205.APR2023
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng đề tài
	  ○ Viết nội dung đề cương
	  ○ Thiết kế poster
	  ○ Làm video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

TẠO VIDEO TỰ ĐỘNG TỪ VĂN BẢN VỚI DIFFUSION MODEL

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

TEXT-TO-VIDEO WITH DIFFUSION MODEL

TÓM TẮT *(Tối đa 400 từ)*

Diffusion Model đã đạt được thành công to lớn trong việc chuyển đổi dữ liệu văn bản thành hình ảnh Text-to-Image (T2I). Điều này mở ra triển vọng vô cùng quan trọng cho tương lai của Text-to-Video (T2V). Với sự đóng góp của Diffusion Model, T2V có thể tiến xa hơn trong việc tạo ra các video tự động từ nội dung văn bản. Hiệu quả của Diffusion Model trong T2I sẽ hỗ trợ cho việc phát triển các mô hình T2V đáng tin cậy và hiệu quả hơn.

Sự kết hợp giữa xử lý ngôn ngữ tự nhiên và thị giác máy tính đã mở ra tiềm năng to lớn cho các ứng dụng T2V trong nhiều lĩnh vực. Từ việc cải thiện trải nghiệm người dùng, tự động tạo nội dung đa phương tiện, hỗ trợ người khuyết tật, đến sử dụng trong giáo dục và truyền thông số, tương lai của Diffusion Model đối với T2V hứa hẹn đem lại nhiều tiện ích và thay đổi cách chúng ta tương tác với dữ liệu văn bản và hình ảnh trong tương lai.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Tạo Video Tự Động từ Văn bản với Diffusion Model là lĩnh vực hứa hẹn trong nghiên cứu Trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên. Thời đại số hóa đòi hỏi việc tạo ra nội dung đa phương tiện động ngày càng quan trọng. Tuy nhiên, việc tạo video thủ công mất nhiều công sức, thời gian và nhân lực. Diffusion Model xuất hiện như giải pháp tiềm năng giúp giảm thiểu khó khăn này và mang lại lợi ích cho giải trí, quảng cáo, giáo dục và truyền thông số.

Một trong những lý do quan trọng cần nghiên cứu là nâng cao trải nghiệm người dùng. Tạo video từ văn bản cải thiện trải nghiệm đa phương tiện, giúp người dùng

tương tác với nội dung một cách sinh động hơn. Diffusion Model cung cấp giải pháp hiệu quả và chất lượng cao để tạo video tự động từ văn bản, thu hút sự chú ý của người xem.

Nghiên cứu còn mang lại tiềm năng tiết kiệm thời gian và chi phí. Tạo video thủ công tốn kém và tốn thời gian, đòi hỏi sự can thiệp của con người. Với Diffusion Model, quá trình này tự động hóa, giảm thiểu thời gian và chi phí cho việc tạo nội dung đa phương tiện.

Diffusion Model đóng góp vào sự phát triển của trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên. Tạo Video Tự Động từ Văn bản yêu cầu kết hợp các kỹ thuật tiên tiến trong hai lĩnh vực này. Nghiên cứu không chỉ mở rộng khả năng ứng dụng của trí tuệ nhân tạo, mà còn cải tiến và phát triển phương pháp xử lý ngôn ngữ tự nhiên và thị giác máy tính. Diffusion Model là cột mốc quan trọng trong tiến bộ công nghệ và định hình tương lai của tạo video tự động từ văn bản.

Triển vọng của nghiên cứu rất hứa hẹn. Diffusion Model mở ra tiềm năng ứng dụng trong giáo dục, truyền thông số và nhiều lĩnh vực khác. Trong giáo dục, T2V cải thiện việc truyền đạt kiến thức và hấp dẫn học sinh qua nội dung giảng dạy động. Trong truyền thông số, Diffusion Model tối ưu hóa chiến lược tiếp thị và quảng cáo với nội dung video chất lượng cao.

Tuy nhiên, nghiên cứu cần được thực hiện cẩn thận. Đảm bảo tính ổn định và hiệu quả của quá trình tạo video đòi hỏi sự chính xác và đáng tin cậy của mô hình Diffusion Model. Thách thức đối mặt là tối ưu hóa hiệu suất và chất lượng của video kết quả, đòi hỏi tinh chỉnh siêu tham số, tối ưu hóa thuật toán và sử dụng dữ liệu đào tạo đa dạng.

Tổng kết lại, nghiên cứu về Tạo Video Tự Động từ Văn bản với Diffusion Model đóng vai trò quan trọng trong cải thiện trải nghiệm người dùng, tiết kiệm thời gian và chi phí, mở rộng khả năng ứng dụng của trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên. Tiềm năng của nghiên cứu này rất lớn và đòi hỏi sự nỗ lực và tập trung của cộng đồng nghiên cứu để đạt được các giải pháp đáng tin cậy và hiệu quả. Tạo Video Tự Động từ Văn bản với Diffusion Model là một hướng đi đầy triển vọng, hứa hẹn đem lại nhiều

lợi ích trong thời đại số hóa và phát triển công nghệ ngày nay.

MỤC TIÊU

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

- Tìm hiểu nghiên cứu liên quan: Mục tiêu này tập trung vào việc tìm hiểu các nghiên cứu, công trình liên quan đến Tạo Video Tự Động từ Văn bản và Diffusion Model. Nghiên cứu sẽ xem xét các phương pháp, mô hình và kỹ thuật đã được đề xuất trong lĩnh vực này và đánh giá hiệu quả của chúng. Tìm hiểu sâu hơn về các tiến bộ và thách thức trong việc tạo video tự động từ văn bản là quan trọng để định hình phạm vi và tiếp cận nghiên cứu một cách hiệu quả.
- Thu thập dữ liệu: Mục tiêu này tập trung vào việc thu thập dữ liệu văn bản và video phù hợp để huấn luyện và đánh giá mô hình Tạo Video Tự Động. Dữ liệu văn bản có thể bao gồm các đoạn văn, mô tả, hoặc script của video. Dữ liệu video phải bao gồm các tập dữ liệu đã được thực hiện thủ công để có thể so sánh và đánh giá hiệu quả của mô hình Diffusion Model so với việc tạo video thủ công.
- Thực nghiệm, chọn độ đo đánh giá và so sánh với việc làm thủ công: Mục tiêu này tập trung vào việc chọn các độ đo đánh giá phù hợp để đánh giá chất lượng và hiệu quả của Tạo Video Tự Động từ Văn bản với Diffusion Model. Độ đo này có thể bao gồm độ tương tự, đánh giá chất lượng hình ảnh và âm thanh, cũng như sự tự nhiên và hấp dẫn của video kết quả. Sau khi chọn độ đo, nghiên cứu sẽ tiến hành so sánh kết quả của mô hình Diffusion Model với việc tạo video thủ công bằng cách sử dụng tập dữ liệu đã thu thập được.

NỘI DUNG VÀ PHƯƠNG PHÁP

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

- Tìm hiểu nghiên cứu liên quan: tiến hành tìm hiểu kỹ lưỡng các công trình nghiên cứu, bài báo, và tài liệu liên quan đến lĩnh vực Tạo Video Tự Động và Diffusion Model. Sự đánh giá và phân tích sẽ tập trung vào các phương pháp và mô hình được sử dụng, từ đó định hình rõ ràng phạm vi của nghiên cứu. Việc

sử dụng các cơ sở dữ liệu khoa học như IEEE Xplore, Google Scholar, và ArXiv sẽ giúp thu thập những công trình uy tín nhất và cập nhật nhất trong lĩnh vực này. Đầu tiên phải tìm hiểu thời gian qua Diffusion Model có những phát triển và cải tiến như thế nào từ các survey[1][2], sau đó tìm hiểu tới nguồn gốc và cải tiến trong T2I của Diffusion Model [3][4] gồm điểm mạnh, yếu, ưu nhược điểm, lý do dẫn tới việc nó bùng nổ như hiện nay. Tiếp theo là việc tìm hiểu tới các nghiên cứu về ứng dụng của Diffusion Model trong T2V từ các đoàn đội lớn như NVIDIA[5], GOOGLE[6], META[7], và các trường đại học có nghiên cứu liên quan [8][9] nhằm biết được phương pháp ứng dụng, phương hướng thực nghiệm, kết quả đạt được, … để có thể rút ra kết luận nên xây dựng giả thuyết nghiên cứu và kiểm chứng thực nghiệm theo phương pháp nào hoặc bằng cách kết hợp các phương pháp trên.

- Thu thập dữ liệu: xây dựng một tập dữ liệu đủ lớn, đa dạng và đại diện để huấn luyện và đánh giá mô hình Tạo Video Tự Động. Dữ liệu văn bản sẽ được thu thập từ các nguồn đáng tin cậy và được tiền xử lý để chuẩn hóa và chuyển đổi thành định dạng phù hợp cho mô hình. Dữ liệu video sẽ bao gồm một tập các video đã được tạo thủ công, cùng với các siêu dữ liệu liên quan như kịch bản và thời gian xuất hiện của các phần tử trong video. Có thể phải xây dựng crawler và tạo danh sách các site để crawl từ các trang có nguồn video phong phú như YouTube, TikTok, Facebook, … và làm sạch lại dữ liệu như việc kiểm tra lại kịch bản, subtitle, thời gian tạo, độ phổ biến, ...
- Thực nghiệm, chọn độ đo đánh giá và so sánh với video thủ công: chạy thực nghiệm và xác định các độ đo đánh giá thích hợp để đo lường chất lượng và hiệu quả của mô hình sau khi huấn luyện. Điều này có thể bao gồm độ tương tự giữa video tự động và video thủ công, sự tương đồng của các phần tử trong video, đánh giá hình ảnh và âm thanh. Đối với độ đo định tính, chúng ta có thể sử dụng các tiêu chí như tự nhiên, sáng tạo, và hấp dẫn. Đồng thời, các độ đo định lượng sẽ được xác định để đánh giá hiệu quả và hiệu suất của mô hình Diffusion Model so với việc tạo video thủ công. Khi đánh giá sẽ thiết lập cả 2

độ đo định tính và định lượng, nhằm tránh trường hợp video được sinh ra bị trùng hoàn toàn với video có trong bộ dữ liệu.

KẾT QUẢ MONG ĐỢI

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

Sau khi hoàn thành quá trình nghiên cứu và thực nghiệm, mong đợi sẽ phát triển thành công một mô hình Tạo Video Tự Động từ Văn bản. Mô hình này sẽ được đánh giá với kết quả đạt được sự tương đồng cao với video tạo thủ công, đảm bảo tính tự nhiên và hấp dẫn của video tự động. Các độ đo đánh giá chất lượng hình ảnh và âm thanh sẽ đạt mức đáng kể, cho thấy khả năng tái tạo chân thực của mô hình. Ngoài ra, tính sáng tạo và độ tương tự giữa video tự động và video thủ công cũng sẽ đạt kết quả ấn tượng.

Việc tự động hóa quá trình tạo video từ văn bản sẽ giúp tiết kiệm thời gian và chi phí, đồng thời cải thiện trải nghiệm người dùng thông qua nội dung đa phương tiện sinh động và hấp dẫn. Mô hình Tạo Video Tự Động từ Văn bản sẽ giúp cho việc sản xuất nội dung đa phương tiện trở nên dễ dàng và hiệu quả hơn, từ đó đáp ứng nhu cầu ngày càng cao của xã hội về nội dung động và đa dạng.

Nếu mô hình đạt được kết quả như mong đợi, điều này sẽ mở ra tiềm năng ứng dụng rộng rãi trong nhiều lĩnh vực như giáo dục, truyền thông số, quảng cáo và giải trí. Sự tiến bộ và thành công của nghiên cứu này sẽ đóng góp quan trọng cho sự phát triển của trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên, định hình tương lai của tạo video tự động từ văn bản. Nó cũng sẽ tạo ra tiềm năng tiết kiệm thời gian và tài nguyên đáng kể cho các tổ chức và cá nhân, đồng thời cải thiện trải nghiệm người dùng trong việc tiếp cận và tương tác với nội dung đa phương tiện.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, Bin Cui: Diffusion Models: A Comprehensive Survey of Methods and Applications. CoRR abs/2209.00796 (2022)

[2] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Mubarak Shah:

Diffusion Models in Vision: A Survey. CoRR abs/2209.04747 (2022)
[3] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, Surya Ganguli: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. CoRR abs/1503.03585 (2015)
[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer: High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022: 10674-10685
[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, Karsten Kreis: Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. CoRR abs/2304.08818 (2023)
[6] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, David J. Fleet: Video Diffusion Models. NeurIPS 2022
[7] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, Yaniv Taigman: Make-A-Video: Text-to-Video Generation without Text-Video Data. ICLR 2023
[8] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, Jie Tang: CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. ICLR 2023
[9] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, Liang Lin: Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models. CoRR abs/2305.13840 (2023)