TẠO VIDEO TỰ ĐỘNG TỪ VĂN BẢN VỚI DIFFUSION MODEL

Huỳnh Khánh Hòa - 220101005

Tóm tắt

- Lớp: CS2205.APR2023
- Link Github: https://github.com/hoahk-uitsdh17/CS2205.APR2023
- Link YouTube video: https://youtu.be/6sDWk28-4k0
- Ảnh + Họ và Tên: Huỳnh Khánh Hòa - 220101005

Giới thiệu

- 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.

Mục tiêu

- 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.

Mục tiêu

- 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.

Mục tiêu

- 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

- Tìm hiểu từ các survey[1][2], nguồn gốc và cải tiến trong T2I của Diffusion Model [3][4], các nghiên cứu liên quan trong T2V NVIDIA[5], GOOGLE[6], META[7], và các trường đại học[8][9]
- 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 bằng định tính và định lượng

Kết quả dự kiến

- 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.

Tài liệu tham khảo

- [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

Tài liệu tham khảo

- [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)