Làm "sống lại" vật thể từ giọng nói tiếng Việt

Sinh viên Nguyễn Văn Anh Tuấn & Phạm Minh Tuấn & Đoàn Minh Trường
Giảng viên dạy PGS. TS Huỳnh Trung Hiếu & Lưu Giang Nam
Hội đồng phản biện PGS. TS Huỳnh Trung Hiếu & TS. Nguyễn Chí Kiên & TS. Bùi Thanh Hùng & Lưu Giang Nam

Tóm tắt nội dung

Nghiên cứu cho thấy rằng hơn một nửa số ngôn ngữ trên thế giới không có dạng viết, điều này khiến cho việc áp dụng nghiên cứu về Xử lý ngôn ngữ tự nhiên cho các ngôn ngữ này gặp hạn chế hơn. Bên cạnh đó, có các ngôn ngữ tuy có dạng viết nhưng số lượng nghiên cứu cho các ngôn ngữ này cũng hạn chế, ví dụ như tiếng Việt. Mặt khác, các nghiên cứu về mạng Thần kinh đối nghịch (GAN) trong bài toán tạo ảnh đang ngày có nhiều nghiên cứu chuyên sâu hơn, nhưng thường tập trung vào dạng viết hoặc ánh xạ từ hình ảnh sang. Trong bài nghiên cứu này, nhóm chúng tôi đề xuất một phương pháp tạo hình ảnh Chim và Hoa từ giọng nói của ngôn ngữ tiếng Việt (tác vụ này gọi là Speech-to-Image Generation - S2IG), tạo tiền đề cho những nghiên cứu về GAN cho giọng nói tiếng Việt sau này. Mô hình thực hiện tác vụ S2IG này được đặt tên là S2IGAN, mô hình kết hợp mạng nhúng giọng nói (Speech Embedding Network - SEN) và mô hình sinh xếp dày đặc được giám sát các mối quan hệ (Relation-supervised densely-stacked generative model - RDG). SEN là một mô hình đa phương thức (multimodal model) học cách phối hợp giọng nói và hình ảnh để đưa ra một thông tin duy nhất. Thông tin này được RDG nhận lấy để tổng hợp hình ảnh có ngữ nghĩa tương ứng với thông tin giọng nói được đưa vào. Kết quả của nghiên cứu này được thực nghiệm trên bộ dữ liệu CUB (ảnh Chim) và Oxford-102 (ảnh Hoa), cả hai bộ dữ liệu này đều thuộc tác vụ Text-to-image Generation (T2IG). Dữ liệu giọng nói tiếng Việt được thu được từ API Text-to-Speech của ViettelAI, dữ liệu chữ viết được tổng hợp dựa trên bản dịch từ Anh sang Việt dùng MTeT của VietAI. Từ thực nghiệm, nhóm tin rằng đây là kết quả khả quan và là nghiên cứu nền cho tác vụ tạo hình ảnh từ giọng nói.

Bố cục bài báo cáo

Bài báo cáo bao gồm 6 phần:

1. Giới thiệu:
   - Giới thiệu về mạng GAN
   - Giới thiệu về S2IG
   - Giới thiệu về những nghiên cứu đã có của S2IG và Text-to-image (T2IG)
2. Phương pháp:
   (a) Mạng nhúng giọng nói (SEN):
   - Trình bày về kiến trúc của mạng nhúng giọng nói kết hợp thông tin của ảnh và giọng nói.
   - Hàm mục tiêu
   (b) Mô hình sinh xếp dày đặc được giám sát các mối quan hệ (RDG): Mạng sinh xếp dày đặc (DG), Giám sát quan hệ (RS), Hàm mục tiêu
3. Thực nghiệm và kết quả
   (a) Dữ liệu: Giới thiệu dữ liệu CUB và Oxford-102
   - Tạo chữ tiếng Việt;
   - Tạo chữ tiếng việt từ chữ tiếng anh từ dữ liệu gốc
   - Tạo dữ liệu giọng nói tiếng Việt
   - Tạo giọng nói tiếng Việt từ chữ tiếng anh
   (b) Cài đặt thực nghiệm: Các thông số, cấu hình cho mô hình
   (c) Kết quả và mục tiêu kết quả: Mục tiêu kết quả; Kết quả
4. Bàn luận và thảo luận
5. Hướng nghiên cứu tương lai
6. Tham khảo

Tài liệu

[1] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, O. Scharenborg, S2IGAN: Speech-to-Image Generation via Adversarial Learning (2020) *arXiv:2005.06968*, doi: 10.48550/arXiv.2005.06968.

[2] X. Wang, T. Qiao, J. Zhu, A. Hanjalic and O. Scharenborg, Generating Images From Spoken Descriptions, in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 850-865, 2021, doi: 10.1109/TASLP.2021.3053391.

[3] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma, W. Gao, Direct Speech-to-image Translation (2020) *arXiv:2004.03413*, doi: 10.48550/arXiv.2004.03413.