|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  | Bài toán tổng hợp tiếng nói là task tạo ra tín hiệu tiếng nói tương ứng với văn bản đầu vào đã cho.  Đầu ra được kỳ vọng nghe tự nhiên nhất có thể (ngữ điệu giống với người nói) và dễ hiểu (phát âm đúng).  Ứng dụng:  + Hỗ trợ người khiếm thị /khuyết tật/ tổn thương thanh quản có thể đọc hiểu văn bản.  + Nhận dạng giọng nói tự động và dịch máy, loại bỏ rào cản giữa những người nói các ngôn ngữ khác nhau.  Framework của bài toán TTS gồm có 2 giai đoạn chính:  GĐ1: Input: văn bản, output: biểu diễn trung gian thường là mel-spectrogram.  GĐ2: input: mel-spectrogram, output: audio dạng sóng  Trong nghiên cứu này nhóm tác giải tập trung vào GĐ2 là tổng hợp âm thanh dạng sóng thô từ biểu diễn trung gian. |
|  |  |
|  | Ở GĐ2 này có nhiều công trình được tiến hành để cải thiện chất lượng tổng hợp âm thanh như:  **Wavenet:** là một mạng nơ ron tích chập tự hồi quy (AR), phương pháp này vượt qua các phương pháp thông thường về chất lượng. Tuy nhiên, nhược điểm của cấu trúc AR làm *chậm trong việc tổng hợp âm thanh có độ phân giải cao theo thời gian*.  Các mô hình Flow-based generative được đề xuất để giải quyết vấn đề này nhờ xử lý song song tăng tốc độ lấy mẫu.  Parallel WaveNet (Oord et al., 2018) cải thiện tốc độ tổng hợp lên 1.000 lần hoặc hơn mà không làm giảm chất lượng.  WaveGlow (Prenger và cộng sự, 2019 tạo ra âm thanh chất lượng cao so với WaveNet. Tuy nhiên, nó đòi hỏi nhiều tham số với hơn 90 layers.  Ngoài ra các mô hình GAN như Mel-GAN, GAN-TTS chiếm ưu thế trong tổng hợp tiếng nói. VD: hiệu quả tham số tốt hơn và thời gian đào tạo tốt hơn.  Mặc dù có những ưu điểm nhưng vẫn có **khoảng cách trong chất lượng mẫu** giữa các mô hình GAN và AR(hoặc flow-based models)  => Nhóm tác giả đề xuất HiFi-GAN, đạt được cả hiệu quả tính toán và chất lượng mẫu cao hơn so với AR hoặc các mô hình flow-based. Vì âm thanh lời nói bao gồm các tín hiệu hình sin với các chu kỳ khác nhau, nên việc lập mô hình các mẫu theo chu kỳ đóng vai trò quan trọng để tạo ra âm thanh lời nói trung thực.  Do đó, nhóm tác giả đề xuất discriminator bao gồm các sub-discriminator, mỗi sub-discriminator gồm một phần chu kỳ cụ thể của waveform thô. Kiến trúc này chính là nền tảng giúp mô hình tổng hợp thành công âm thanh giọng nói trung thực. |
|  |  |
|  | HiFi-GAN bao gồm một generator và hai discriminators: multi-scale và multi-period discriminators. Generator và discriminator được huấn luyện đối kháng, cùng với hai loss bổ sung để cải thiện độ ổn định huấn luyện và hiệu suất mô hình. |
| *Hình 1. Generator lấy mẫu mel-spectrogram lên đến |ku| lần để khớp với độ phân giải tạm thời của dạng sóng thô. Mô-đun MRF thêm các tính năng từ |kr| các khối còn lại kernel size và tốc độ giãn nở khác nhau. Cuối cùng, khối dư n-th với kích thước kernel kr[n] và tốc độ giãn nở Dr[n] trong mô-đun MRF được mô tả.* | Generator là một mạng fully CNN. Sử dụng một mel-spectrogram làm đầu vào và lấy mẫu thông qua tích chập chuyển vị cho đến khi độ dài của chuỗi đầu ra khớp với độ phân giải tạm thời của dạng sóng thô. Mỗi tích chập chuyển vị được theo sau bởi một mô đun hợp nhất trường (Multi-receptive field fusion - MRF). mô-đun này quan sát song song các mẫu có độ dài khác nhau. Cụ thể, mô-đun MRF trả về tổng đầu ra từ nhiều Residual Block. Các kích thước kernel và tốc độ giãn nở khác nhau được chọn cho từng khối còn lại để tạo thành các mẫu trường tiếp nhận đa dạng. |
|  | Mục tiêu của Discriminator là phân lớp xem waveform của generator tạo ra là real/fake nếu mà đc nhận là real thì xem như âm thanh tạo ra bởi generator tốt, càng gần với tiếng nói tự nhiên.  Có 2 loại discriminator:  **Multi-Period Discriminator**: MPD là sự kết hợp của các sub-discriminators, mỗi bộ chỉ chấp nhận các mẫu âm thanh đầu vào có khoảng cách đều nhau;  không gian được đưa ra như là khoảng thời gian p. Các sub-discriminators được thiết kế để bắt các cấu trúc ẩn khác nhau bằng cách xem xét các phần khác nhau của âm thanh đầu vào. ~~Đặt các khoảng thời gian thành [2, 3, 5, 7, 11] để tránh trùng lặp nhiều nhất có thể.~~  Hoạt động của MPD được hiển thị trong Hình phía dưới, trước tiên reshape audio thô 1D có độ dài T thành dữ liệu 2D có chiều cao T /p và chiều rộng p , sau đó tích chập 2D trên dữ liệu reshape. Điểm khác so với các nghiên cứu trước là chuyển sang 2D thay vì lấy mẫu trực tiếp trên tín hiệu audio.  ~~Trong mỗi lớp tích chập của MPD, giới hạn kích thước kernel ở trục chiều rộng là 1 để xử lý các mẫu định kỳ một cách độc lập. Mỗi sub-discriminator là một chồng các lớp tích chập có bước kích hoạt đơn vị tuyến tính được chỉnh lưu (leaky ReLU). Sau đó, chuẩn hóa weight (Salimans và Kingma, 2016) được áp dụng cho MPD. Bằng cách reshape âm thanh đầu vào thành dữ liệu 2D thay vì lấy mẫu tín hiệu trên audio.~~ |
|  | **Multi-Scale Discriminator**: Bởi vì mỗisub-discriminator trong MPD chỉ chấp nhận các mẫu rời rạc, nên thêm MSD để đánh giá chuỗi âm thanh một cách liên tục. Kiến trúc của MSD được rút ra từ kiến trúc của MelGAN (Kumar et al., 2019). MSD là sự kết hợp của ba sub-discriminators hoạt động trên các scale đầu vào khác nhau: raw audio , audio tổng hợp trung bình ×2 và audio tổng hợp trung bình ×4, như thể hiện trong Hình 2a.  Mỗi sub-discriminators trong MSD là một chồng các lớp tích chập được sắp xếp và nhóm lại với kích hoạt leaky ReLU.  Kích thước của discriminator được tăng lên bằng cách giảm stride và thêm nhiều layer. Weight normalization được áp dụng ngoại trừ bộ phân biệt đối xử phụ đầu tiên hoạt động trên âm thanh thô. Thay vào đó, chuẩn hóa phổ (Miyato et al., 2018) được áp dụng và ổn định quá trình đào tạo như đã báo cáo. |
|  | **GAN Loss:**  Discriminator được đào tạo để phân loại các mẫu grouth truth thành 1 và các mẫu được tổng hợp từ generator thành 0. Generator được huấn luyện để đánh lừa bộ discriminator bằng cách cập nhật chất lượng mẫu để được phân loại thành giá trị gần bằng 1. GAN losses cho generator G và Discriminator D được định nghĩa:  trong đó x biểu thị ground truth audio và s biểu thị điều kiện đầu vào, mel spectrogram của ground truth audio. |
|  | **Mel-Spectrogram Loss** Ngoài GAN loss, thêm mel-spectrogram loss để cải thiện hiệu quả huấn luyên của trình tạo và độ trung thực của audio được tạo ra. mel spectrogram loss là khoảng cách L1 giữa mel-spectrogram của waveform do generator tổng hợp và của ground truth waveform . |
|  | **Feature Matching Loss** là một độ đo sự giống nhau bằng sự khác biệt về các tính năng của discriminator giữa ground truth sample và mẫu được tạo.  trong đó T biểu thị số lượng layer trong discriminator; Di và Ni biểu thị các đặc trưng và số lượng đặc trưng trong lớp thứ i của discriminator tương ứng. |
|  |  |
|  |  |
|  | sử dụng bộ dữ liệu LJSpeech (Ito, 2017) bao gồm 13.100 đoạn âm thanh ngắn của một người nói với tổng thời lượng khoảng 24 giờ. Định dạng âm thanh là PCM 16 bit với tốc độ lấy mẫu là 22 kHz;  HiFi-GAN được so sánh với các mô hình tốt nhất có sẵn công khai: WaveNet, MelGAN và WaveGlow. sử dụng pretrain weight cung cấp cho tất cả các mô hình.  Để đánh giá tính tổng quát của HiFi-GAN đối với mel-spectrogram inversion của những người nói không nhìn thấy, sử dụng bộ dữ liệu nhiều người nói VCTK (Veaux et al., 2017), bao gồm khoảng 44.200 đoạn âm thanh ngắn được phát ra bởi 109 người nói tiếng Anh bản địa với nhiều điểm nhấn khác nhau. Tổng thời lượng của các đoạn âm thanh là khoảng 44 giờ. Audio format là PCM 16 bit với tốc độ lấy mẫu là 44 kHz. Giảm tốc độ lấy mẫu xuống 22 kHz. Nhóm tác giả đã chọn ngẫu nhiên 9 speaker và loại trừ tất cả các đoạn âm thanh của họ khỏi tập huấn luyện. Sau đó, đã đào tạo MoL WaveNet, WaveGlow và MelGAN với cùng data setting; tất cả các mô hình đã được đào tạo cho đến 2,5 triệu step.  Để đánh giá chất lượng audio, thu thập các bài MOS test 5 scale của Amazon Mechanical Turk. Điểm MOS được ghi lại  Để xác nhận sự đánh đổi giữa hiệu quả tổng hợp và chất lượng mẫu, tiến hành các thử nghiệm dựa trên ba biến thể của generator, V1, V2 và V3 trong khi vẫn duy trì cấu hình discriminator giống nhau. |
|  |  |
|  |  |
|  | ~~Để đánh giá hiệu suất của các mô hình của chúng tôi về cả chất lượng và tốc độ, chúng tôi đã thực hiện thử nghiệm MOS để đảo ngược phổ và đo tốc độ. Đối với bài kiểm tra MOS, chúng tôi đã chọn ngẫu nhiên 50 cách phát biểu từ bộ dữ liệu LJSpeech và sử dụng biểu đồ phổ chân lý cơ bản của các cách phát biểu đã bị loại khỏi quá trình đào tạo làm điều kiện đầu vào.~~  Để so sánh chất lượng audio, tốc độ tổng hợp và kích thước mô hình, các kết quả được tổng hợp và trình bày trong Bảng 1. Đáng chú ý, tất cả các biến thể của HiFi-GAN đều đạt điểm cao hơn các mẫu khác. V1 có thông số 13,92M và đạt MOS cao nhất với khoảng cách 0,09 so với groundtruth audio;   * âm thanh tổng hợp gần như không thể phân biệt được với giọng nói của con người.   Về tốc độ tổng hợp, V1 nhanh hơn WaveGlow và MoL WaveNet. V2 cũng thể hiện chất lượng tương tự như con người với MOS là 4,23 trong khi giảm đáng kể yêu cầu bộ nhớ và thời gian suy luận so với V 1. Nó chỉ yêu cầu 0,92M tham số. Mặc dù có MOS thấp nhất trong số các kiểu máy của chúng tôi, V3 có thể tổng hợp giọng nói nhanh hơn 13,44 lần so với thời gian thực trên CPU và nhanh hơn 1.186 lần so với thời gian thực trên một GPU V100 trong khi vẫn cho thấy chất lượng cảm nhận tương tự với MoL WaveNet. Vì V3 tổng hợp lời nói trên CPU nó có thể rất phù hợp cho các ứng dụng trên thiết bị. |
|  | Nhóm tác giả đã thực hiện một nghiên cứu cắt bỏ MPD, MRF và mel-spectrogram để xác minh ảnh hưởng của từng thành phần HiFi-GAN đối với chất lượng của âm thanh tổng hợp. V3 có công suất biểu đạt nhỏ nhất trong số ba biến thể của bộ tạo đã được sử dụng làm bộ tạo cho nghiên cứu cắt bỏ và các tham số mạng được cập nhật lên tới 500 nghìn bước cho mỗi cấu hình.  Kết quả đánh giá MOS được thể hiện trong Bảng 2, cho thấy cả ba thành phần đều góp phần vào hiệu suất. Việc loại bỏ MPD làm giảm đáng kể chất lượng cảm nhận, trong khi sự vắng mặt của MSD cho thấy sự xuống cấp tương đối nhỏ nhưng đáng chú ý. Để điều tra ảnh hưởng của MRF, một khối còn lại có trường tiếp nhận rộng nhất được giữ lại trong mỗi mô-đun MRF. Kết quả cũng tệ hơn so với đường cơ sở. Thử nghiệm về suy hao mel-spectrogram cho thấy nó giúp cải thiện chất lượng và chúng tôi quan sát thấy rằng chất lượng cải thiện ổn định hơn khi áp dụng suy hao.  Để xác minh tác dụng của MPD trong cài đặt của các kiểu GAN khác, chúng tôi đã giới thiệu MPD trong MelGAN. MelGAN được đào tạo với MPD vượt trội so với bản gốc với khoảng cách 0,47 MOS, điều này cho thấy sự cải thiện đáng kể về mặt thống kê.  Thực nghiệm với các chu kỳ lũy thừa bằng 2 để xác minh ảnh hưởng của các chu kỳ được đặt thành các số nguyên tố. Mặc dù giai đoạn 2 cho phép các tín hiệu được xử lý chặt chẽ, nhưng nó dẫn đến sự xuống cấp đáng kể về mặt thống kê với mức chênh lệch 0,20 MOS so với baseline |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |