**Nội dung của Nghiên cứu:**

1. **Text-to-Music:** Nghiên cứu tập trung vào việc tạo ra âm nhạc dựa trên mô tả văn bản, như "bài hát rock thập kỷ 90 với giai điệu guitar."
2. **Khó khăn trong Tạo Âm Nhạc:** Quá trình tạo ra âm nhạc đầy thách thức vì cần mô hình hóa chuỗi dài, sử dụng toàn bộ phổ tần số, và xử lý cấu trúc phức tạp của âm nhạc.

**Các Tiến Bộ Gần Đây và Phương Pháp:**

1. **Học Biểu Diễn Âm Thanh:** Sử dụng tiến bộ trong học biểu diễn âm thanh tự học, mô hình tuần tự, và tổng hợp âm thanh để phát triển các mô hình tạo âm nhạc.
2. **Biểu Diễn Âm Thanh Dưới Dạng Nhiều Luồng Mã Thông Báo:** Đề xuất biểu diễn âm thanh dưới dạng nhiều luồng mã thông báo rời rạc để tăng cường chất lượng âm thanh và hiệu suất.

**Công Trình Nghiên Cứu Mới:**

1. **Mô Hình MUSIGEN:** Giới thiệu MUSIGEN, một mô hình đơn giản và dễ điều khiển để tạo ra âm nhạc chất lượng cao dựa trên mô tả văn bản.
2. **Khung Công Việc Tổng Quát:** Đề xuất một khung công việc tổng quát để mô hình hóa nhiều luồng âm thanh song song của các đơn vị âm thanh.
3. **Đánh Giá và So Sánh:** Thực hiện đánh giá chi tiết về MUSIGEN, và chỉ ra rằng phương pháp đề xuất vượt trội so với các phương pháp cơ bản khác.

**Đóng Góp của Nghiên Cứu:**

1. **Mô Hình Đơn Giản và Hiệu Quả:** Giới thiệu mô hình tạo âm nhạc chất lượng cao với tần số 32 kHz thông qua một chiến lược xen kẽ mã code hiệu quả.
2. **Tạo Âm Dựa Trên Văn Bản và Giai Điệu:** Thực hiện tạo âm dựa trên cả văn bản và giai điệu, đồng thời đảm bảo âm thanh phù hợp và trung thành với thông tin điều kiện văn bản.
3. **Đánh Giá Mục Tiêu và Đánh Giá của Con Người:** Cung cấp đánh giá mục tiêu và đánh giá của con người về các lựa chọn thiết kế quan trọng trong phương pháp của nghiên cứu.

**Phương pháp:**

1. **Mô Hình MUSICGEN:**
   * **Loại Mô Hình:** Mô hình decoder dựa trên transformer autoregressive.
   * **Cơ Sở Lý Thuyết:** Sử dụng transformer autoregressive theo kiến trúc được giới thiệu trong Vaswani et al., 2017.
   * **Điều Kiện Đầu Vào:** Mô hình được điều kiện dựa trên biểu diễn văn bản hoặc giai điệu.
2. **EnCodec:**
   * **Mã Hóa Âm Thanh:** Sử dụng EnCodec, một mô hình tự mã hóa tích chập, với lượng tử hóa bằng Residual Vector Quantization (RVQ).
   * **Lượng Tử Hóa:** Áp dụng lượng tử hóa để tạo ra nhiều luồng mã thông báo rời rạc từ codebook đã học.
3. **Audio Tokenization:**
   * **Quy Trình Mã Hóa âm thanh:** Sử dụng EnCodec để chuyển đổi âm thanh thành tensor liên tục với tốc độ khung nhỏ hơn.
   * **Lượng Tử Hóa và Codebook:** Sử dụng Residual Vector Quantization (RVQ) để tạo ra nhiều luồng mã thông báo rời rạc từ codebook đã học.
4. **Codebook Interleaving Patterns:**
   * **Phân Rã Tự Hồi Quy:** Sử dụng mô hình tự hồi quy và mẫu mã thông báo codebook để phân rã tự hồi quy phẳng chính xác.
5. **Model Conditioning:**
   * **Điều Kiện Bằng Văn Bản và Giai Điệu:** Sử dụng tensor điều kiện được tính toán từ mô tả văn bản hoặc giai điệu.
6. **Model Architecture:**
   * **Projection và Positional Embedding:** Sử dụng projection từ codebook và positional embedding trong mô hình để mã hóa bước hiện tại.
   * **Transformer Decoder:** Sử dụng trình giải mã transformer với các khối tự-chú ý và chú ý chéo được cung cấp bởi tín hiệu điều kiện.
   * **Logits Prediction:** Dự đoán logits cho các giá trị của Q tại các chỉ số được chỉ định.

**Xử lý Dữ Liệu:**

* **Tokenization:** Sử dụng EnCodec cho quá trình mã hóa âm thanh và tạo các token rời rạc từ codebook đã học.
* **Conditioning:** Sử dụng mô hình tự học cho việc điều kiện dựa trên văn bản hoặc giai điệu.

**Mục Tiêu và Phân Tích:**

* **Mục Tiêu:** Tạo ra âm nhạc chất lượng cao có thể điều khiển dựa trên mô tả văn bản hoặc giai điệu.
* **Phân Tích Mô Hình:** Thử nghiệm và phân tích các phương pháp mã hóa, điều kiện, và kiến trúc để đánh giá hiệu suất.

**Kết Luận:** Nghiên cứu giới thiệu mô hình MUSICGEN, kết hợp transformer autoregressive, EnCodec, và các chiến lược điều kiện để tạo ra âm nhạc có chất lượng cao và có thể điều khiển dựa trên văn bản hoặc giai điệu. Các phương pháp được sử dụng như lượng tử hóa, mô hình tự hồi quy, và điều kiện đầu vào từ văn bản hoặc giai điệu. Thêm vào đó, các mô hình được đánh giá thông qua đánh giá mục tiêu và đánh giá của con người.

**Thiết Lập Thử Nghiệm (Experiment Setup):**

1. **Mô Hình và Siêu Tham Số (Models and Hyperparameters):**
   * **Mô Hình Mã Hóa Âm Thanh (Audio Tokenization Model):**
     + Sử dụng EnCodec với năm lớp không gian cho âm thanh đơn âm 32 kHz.
     + Bước nhảy: 640, tạo ra tốc độ khung hình là 50 Hz.
     + Kích thước ẩn ban đầu: 64, tăng gấp đôi ở mỗi lớp.
     + Sử dụng RVQ (Residual Vector Quantization) với bốn bộ lượng tử và kích thước bảng mã là 2048.
     + Huấn luyện trên đoạn âm thanh một giây được cắt ngẫu nhiên từ chuỗi âm thanh.
   * **Mô Hình Transformer:**
     + Huấn luyện với các kích thước khác nhau: 300 triệu, 1.5 tỷ, và 3.3 tỷ tham số.
     + Sử dụng Flash attention từ gói xFormers để cải thiện tốc độ và sử dụng bộ nhớ tốt hơn với các chuỗi dài.
     + Sử dụng AdamW và D-Adaptation để tối ưu hóa hiệu suất và độ ổn định.
     + Huấn luyện trên đoạn âm thanh kéo dài 30 giây được lấy mẫu ngẫu nhiên từ bản gốc.
   * **Xử Lý Văn Bản (Text Processing):**
     + Sử dụng phương pháp "text-normalization" để chuẩn hóa văn bản, loại bỏ từ dừng và lemmatize phần còn lại của văn bản.
     + Thử nghiệm việc ghép các nhãn chú thích bổ sung vào mô tả văn bản (phương pháp "condition-merging").
     + Áp dụng "word dropout" làm chiến lược tăng cường văn bản.
   * **Mô Hình Mã Hóa Văn Bản (Text Encoding Models):**
     + Sử dụng mô hình mã hóa văn bản T5, FLAN-T5, và CLAP để điều kiện văn bản.
   * **Mô Hình Điều Kiện Giai Điệu (Melody Conditioning Model):**
     + Tính toán chromagram với kích thước cửa sổ là 214 và bước nhảy là 212.
     + Sử dụng phương pháp lấy mẫu từ logits của mô hình, với xác suất loại bỏ điều kiện và tỷ lệ hướng dẫn cụ thể.
2. **Patterns và Điều Kiện Codebook:**
   * Sử dụng mô hình mã hóa văn bản T5 để điều kiện văn bản và thử nghiệm với các mô hình mã hóa văn bản khác như FLAN-T5 và CLAP.
   * Đối với điều kiện giai điệu, tính toán chromagram và sử dụng chiến lược lấy mẫu từ logits của mô hình.
3. **Chiến Lược Tiền Xử Lý Văn Bản (Text Processing Strategies):**
   * Thử nghiệm chiến lược "condition-merging" với xác suất là 0.25 và áp dụng "text description dropout" với xác suất là 0.5.
   * Sử dụng "word dropout" với xác suất là 0.3 trên văn bản kết quả.

**Kết Luận:**

* Sử dụng mô hình mã hóa âm thanh EnCodec và mô hình transformer với các siêu tham số khác nhau.
* Tiền xử lý văn bản bao gồm chuẩn hóa, loại bỏ từ dừng, lemmatize, và các chiến lược tăng cường văn bản.
* Điều kiện mô hình dựa trên mô tả văn bản, nhãn chú thích âm nhạc bổ sung, và giai điệu.

**Bộ Dữ Liệu (Datasets):**

1. **Bộ Dữ Liệu Huấn Luyện (Training Datasets):**
   * Sử dụng tổng cộng 20.000 giờ âm nhạc được cấp phép.
   * Bao gồm một bộ dữ liệu nội bộ với 10.000 bản nhạc chất lượng cao.
   * Bộ sưu tập âm nhạc từ ShutterStock và Pond5 với 25.000 và 365.000 bản nhạc, chứa thông tin mô tả văn bản như thể loại, BPM, và từ khóa.
   * Tất cả đều có tần số lấy mẫu 32 kHz, và có chuyển đổi âm thanh thành đơn âm nếu cần thiết.
2. **Bộ Dữ Liệu Đánh Giá (Evaluation Datasets):**
   * Sử dụng bộ kiểm tra MusicCaps, bao gồm 5.5K mẫu (mỗi mẫu dài 10 giây) chuẩn bị bởi các nhạc sĩ chuyên gia.
   * Tập con 1K cân bằng theo thể loại để đánh giá âm nhạc từ văn bản.
   * Đánh giá trên tập không cân bằng và lấy mẫu các ví dụ từ tập cân bằng theo thể loại để đánh giá các yếu tố như giai điệu và sự sáng tạo.

**Bảng So Sánh (Comparison Table):**

* So sánh các chỉ số khách quan và chủ quan của MUSICGEN với một số cơ sở khác.
* Báo cáo điểm trung bình và khoảng tin cậy CI95.
* Các mô hình khác bao gồm Mousai (đào tạo trên cùng tập dữ liệu), MusicLM (sử dụng API công khai), và Noise2Music.
* Báo cáo FAD ban đầu trên MusicCaps cho Noise2Music và MusicLM.
* "MUSICGEN w. random melody" chỉ đến MUSICGEN được đào tạo với chromagram và văn bản, với việc lấy mẫu chromagram ngẫu nhiên từ một tập dữ liệu giữ lại trong quá trình đánh giá.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

**Đánh Giá (Evaluation):**

1. **So Sánh Cơ Sở (Baseline Comparison):**
   * So sánh với hai mô hình cơ sở (Riffusion và Mousai) cho việc tạo âm nhạc từ văn bản.
   * MUSICGEN được đào tạo trên tập dữ liệu riêng để đảm bảo sự so sánh công bằng.
   * Ngoài ra, so sánh với MusicLM và Noise2Music nếu có thể.
2. **Chỉ Số Đánh Giá (Evaluation Metrics):**
   * **Chỉ Số Mục Tiêu (Objective Metrics):**
     + **Khoảng Cách Âm Thanhh Fréchet (FAD):** Để đánh giá sự hợp lý của âm thanh được tạo ra. Điểm FAD thấp là tốt.
     + **Độ Lệch Kullback-Leiber (KL):** Đo lường sự giống nhau giữa nhãn của âm nhạc gốc và âm nhạc được tạo ra.
     + **Điểm CLAP:** Đo sự phù hợp giữa mô tả bài hát và âm thanh được tạo ra.
   * **Chỉ Số Chủ Quan (Subjective Metrics):**
     + **Đánh Giá Chất Lượng Tổng Thể:** Đánh giá chất lượng tổng thể của âm thanh được tạo ra.
     + **Đánh Giá Sự Liên Quan Đến Nội Dung Văn Bản:** Đánh giá mức độ mà âm thanh liên quan đến nội dung văn bản của nó.
3. **Bảng 2:**
   * Báo cáo độ tương đồng cosine giữa giai điệu tham chiếu và giai điệu được tạo ra (SIM.).
   * Chứa các chỉ số chủ quan bao gồm sự phù hợp với giai điệu (MEL.).
   * Tất cả kết quả được báo cáo cho MUSICGEN 1.5B.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Mô tả được tạo tự động

**Kết Quả (Results):**

1. **So Sánh với Cơ Sở (Comparison with Baselines):**
   * Bảng 1 trình bày sự so sánh giữa phương pháp đề xuất (MUSICGEN) và các phương pháp Mousai, Riffusion, MusicLM và Noise2Music.
   * Đánh giá FAD trên MusicCaps, với Noise2Music cho kết quả tốt nhất, tiếp theo là MUSICGEN được huấn luyện với điều kiện văn bản.
   * Kết quả cho thấy MUSICGEN hoạt động tốt hơn so với các phương pháp cơ sở, cả về chất lượng âm thanh và sự tuân thủ với mô tả văn bản.
2. **Đánh Giá Giai Điệu (Melody Evaluation):**
   * Đánh giá MUSICGEN được điều kiện chung trên biểu diễn văn bản và giai điệu.
   * Chỉ số mới được giới thiệu: độ tương đồng cosine của chroma, đo lường sự tương đồng trung bình giữa các khung tương ứng.
   * Kết quả cho thấy MUSICGEN được đào tạo với điều kiện chromagram thành công trong việc tạo ra âm nhạc theo giai điệu đã cho.
   * MUSICGEN cho phép kiểm soát tốt hơn về đầu ra được tạo ra và có tính ổn định khi loại bỏ chroma trong quá trình suy luận.
3. **Mở Rộng Mẫu Mã Âm Thanh (Extension to Stereo Sound):**
   * Bài viết đề cập đến mở rộng mẫu mã âm thanh cho việc tạo ra âm thanh stereo.
   * Có thể đề cập đến chi tiết cụ thể về cách mô hình được mở rộng và cách nó ảnh hưởng đến chất lượng âm thanh.
4. **Tác Động của Các Yếu Tố Cụ Thể (Impact of Specific Factors):**
   * Nghiên cứu về tác động của các yếu tố cụ thể trong việc tạo ra âm thanh từ văn bản.
   * Có thể đề cập đến các yếu tố như loại biểu diễn văn bản, kích thước mô hình, hoặc các tham số huấn luyện có ảnh hưởng đến kết quả.

**Tổng Kết (Conclusion):**

* Bài viết tổng kết kết quả của phương pháp đề xuất và nhấn mạnh các điểm mạnh so với các nghiên cứu trước đó.
* Cung cấp cái nhìn tổng quan về hiệu suất của MUSICGEN trong việc tạo âm nhạc từ văn bản và đánh giá theo các tiêu chí khác nhau.

Hình 2 trong bài viết là một biểu đồ minh họa về Codebook stereo và kết quả ghi nhớ. Hình này chia thành hai phần: bên trái là Codebook stereo, và bên phải là kết quả ghi nhớ.

1. **Codebook stereo (Bên Trái):**
   * Hiển thị một tập hợp các mẫu âm thanh, được gọi là Codebook stereo.
   * Các mẫu âm thanh này được sử dụng trong mô hình stereo để tạo ra âm thanh với hiệu ứng âm thanh stereo.
2. **Kết Quả Ghi Nhớ (Bên Phải):**
   * Đo lường khả năng ghi nhớ của mô hình trên các mẫu âm thanh.
   * Có thể thấy mức độ ghi nhớ của mô hình dựa trên so sánh giữa âm thanh được tạo ra và Codebook stereo.

Hình này có thể chứa thông tin chi tiết về cách mô hình xử lý và tái tạo âm thanh stereo, cũng như cách khả năng ghi nhớ của mô hình được đánh giá. Nó cung cấp một cách trực quan để hiểu về cách mô hình làm việc và hiệu suất của nó trong việc tạo ra âm thanh stereo từ Codebook stereo.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Nhiều màu sắc

Mô tả được tạo tự động

(a) Hình ảnh hóa các mẫu codebook cho các mô hình stereo với hai cách xen kẽ có thể của codebook trái (màu xanh) và codebook phải (màu đỏ):

* Hình ảnh này có thể hiển thị các mẫu codebook cho các mô hình stereo.
* Màu xanh thường biểu thị codebook hoặc mẫu cho kênh âm thanh trái, trong khi màu đỏ có thể biểu thị codebook hoặc mẫu cho kênh âm thanh phải.
* Có thể có sự xen kẽ giữa các mẫu để thể hiện cách chúng được sử dụng xen kẽ để tạo ra âm thanh stereo.
* Đối với mỗi chỉ số codebook cụ thể, có thể hiển thị mẫu "stereo delay" và "stereo partial delay" để thể hiện độ trễ khác nhau giữa kênh trái và kênh phải.

Ảnh có chứa văn bản, hàng, ảnh chụp màn hình, Sơ đồ

Mô tả được tạo tự động

(b) Kết quả ghi nhớ cho các mã codebook đầu tiên trong việc tạo ra âm thanh trong 5 giây, xem xét sự khớp chính xác (đường liền) và sự khớp một phần 80% (đường đứt) khi được yêu cầu với các đoạn trích có độ dài khác nhau từ tập dữ liệu huấn luyện:

* Đồ thị này có thể hiển thị kết quả ghi nhớ của các mã codebook đầu tiên khi tạo âm thanh.
* Có thể sử dụng đường liền để biểu thị sự khớp chính xác và đường đứt để biểu thị sự khớp một phần 80%.
* Điều này có thể giúp đánh giá khả năng của mô hình ghi nhớ và tái tạo âm thanh stereo dựa trên các mã codebook.