Generate Image from Audio

Alessio Salzano
Università degli Studi di Salerno
Curriculum Sicurezza Informatica
Salerno, Italy
a.salzano32@studenti.unisa.it

Abstract

Questo report presenta un sistema innovativo di generazione di immagini in tempo reale da testo, progettato per applicazioni dal vivo come performance musicali. La pipeline sviluppata utilizza modelli avanzati di intelligenza artificiale. Viene utilizzato il modello Whisper per la trascrizione del testo, il modello T5 per sintetizzare il testo e un modello di Stable Diffusion per la generazione di immagini. I test effettuati su brani celebri dimostrano una coerenza visiva notevole rispetto ai testi delle canzoni. Questo progetto rappresenta un passo avanti nel campo della visualizzazione real-time per esibizioni dal vivo, con possibili sviluppi futuri verso l'integrazione di componenti musicali per una maggiore armonia visiva e sonora.

1 Introduzione

Negli ultimi anni, il ruolo visivo nella musica dal vivo ha acquisito una sempre maggiore rilevanza, contribuendo a migliorare le performance artistiche e musicali. Vi è mai capitato di ascoltare qualcuno che descrive una scena in un discorso o in un audiolibro, immaginandola nella vostra mente? Utilizzando l'intelligenza artificiale, è ora possibile trasformare automaticamente le parole pronunciate in immagini, trascrivendo l'audio e generando rappresentazioni visive. Il presente progetto propone un sistema innovativo per la generazione di immagini in tempo reale, sincronizzate con le esecuzioni musicali dal vivo. L'obiettivo principale è fornire un supporto visivo dinamico e coerente con i testi cantati, proiettando le immagini su un ledwall durante le performance. Questo studio illustra le componenti principali della pipeline sviluppata e i risultati ottenuti attraverso test su canzoni selezionate, offrendo spunti per possibili futuri sviluppi e miglioramenti. Il codice del progetto è disponibile al seguente link

La struttura del presente report viene riportata qui di seguito:

- Nella sezione 2, si analizza l'attuale stato dell'arte in ambito Music to Image Generation.
- Nella sezione 3, si analizza il metodo proposto per generare le immagini da proiettare durante l'esibizione musicale, descrivendo i vari modelli utilizzati.
- Nella sezione 4, si analizzano i risultati ottenuti, mostrando qualche esempio.
- Nella sezione 5, si è discusso l'intero processo traendo le opportune conclusioni, discutendo le limitazioni e proponendo alcuni sviluppi futuri.

2 STATO DELL'ARTE

Molti degli studi presenti in letteratura, in ambito della generazione di immagini a partire dalla musica si concentrano sull'analisi di due tipi di input. Questi studi mirano a esplorare come le caratteristiche musicali e visive possano essere correlate per generare immagini coerenti con la musica. In Qiu et al. (2018)[1], vengono prese in input la musica e le immagini, ed estratte in modo separato determinate caratteristiche. Le caratteristiche estratte includono i ritmi musicali, le tonalità, e le texture visive delle immagini. Attraverso l'uso di reti neurali convoluzionali (CNN) e ricorrenti (RNN), il modello impara la correlazione tra le caratteristiche visive e musicali e, partendo da questa correlazione, genera nuove immagini che sono coerenti con la musica in input. Di contro, Iadodong Tan e Mathis Antony[2] hanno affrontato il problema inverso: data un'immagine, il

loro modello estrae determinate caratteristiche visive, le analizza e genera musica che rispecchia tali caratteristiche. Ad esempio, questa tecnica può essere utilizzata per comprendere quali caratteristiche visive influenzano specifiche qualità musicali e verificare se l'immagine generata da un determinato modello risulta coerente con il brano musicale, andando quindi ad utilizzare questo approccio come metrica di comparazione per testare la coerenza delle immagini generate a partire dalla musica. Nello studio di Meng Yang et al.[3], viene sviluppato un sistema per la generazione di immagini in tempo reale utilizzando dati MIDI provenienti da performance musicali. Il sistema analizza i messaggi MIDI, che contengono informazioni dettagliate su note, intensità e timing. Attraverso l'uso di modelli di intelligenza artificiale generativa, come i modelli di linguaggio naturale (LLM), il sistema interpreta le informazioni estratte, generando caratteristiche emotive e successivamente, genera immagini visive in tempo reale che rappresentano le caratteristiche emotive che sono state ottenute, offrendo anche una sensazione visiva alla performance musicale.

3 Metodologia

Il sistema si basa sull'utilizzo di modelli di generazione di immagini e trascrizione del testo ottimizzati per prestazioni in tempo reale. La pipeline prevede le seguenti fasi:

- Trascrizione del testo.
- Sintesi del Testo.
- Generazione delle Immagini.

3.1 Trascrizione del testo

Il nostro viaggio nella trasformazione da audio ad immagine inizia con **Whisper**[4], un sistema di riconoscimento vocale automatico (ASR) addestrato su 680.000 ore di dati supervisionati multilingue e multitask raccolti dal web. Whisper ASR svolge un ruolo fondamentale nel nostro progetto. È la fase iniziale in cui l'audio viene trasformato in testo. Questo è fondamentale perché non possiamo trasformare direttamente le onde sonore in immagini; abbiamo bisogno di un intermediario, e Whisper svolge questo ruolo in modo eccezionale. Assicura che le parole pronunciate nell'audio siano accuratamente rappresentate in forma scritta.

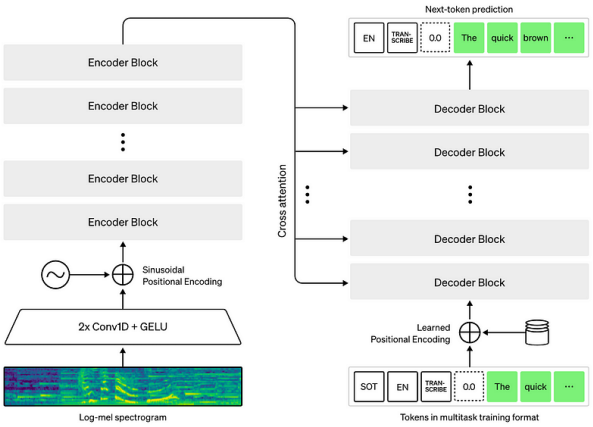

Figure 1: Whisper Architecture.

3.2 Sintesi del Testo

Dopo la trascrizione del testo tramite Whisper, il passo successivo consiste nella sintesi del contenuto trascritto. Questa fase è cruciale per diversi motivi. In primo luogo, il modello di generazione delle immagini successivo non può gestire un'eccessiva quantità di input testuale. Pertanto, è necessario ridurre il testo a frasi rappresentative che preservino il significato originale.

In secondo luogo, sintetizzare il testo aiuta a mantenere la generalità del contenuto, evitando dettagli superflui che potrebbero compromettere la coerenza e l'efficacia delle immagini generate. Infine, un testo conciso consente di ottenere risultati più accurati nella generazione delle immagini, poiché permette al modello di concentrarsi sui concetti chiave, facilitando così l'associazione tra testo e rappresentazione visiva.

Per questo scopo, utilizziamo il modello T5-small, una variante del T5 - Text-to-Text-Transformer[5]. Questo modello riassume il testo trascritto in modo efficace, estraendo le frasi più significative. Grazie a questa sintesi, il contenuto visivo può essere allineato con il testo lirico, migliorando la coerenza e la rilevanza delle immagini generate.

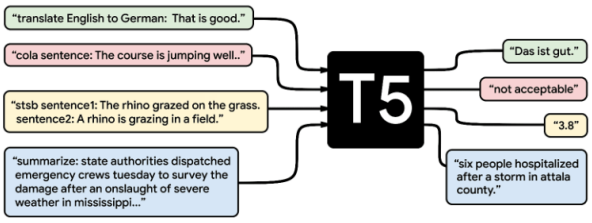

Figure 2: T5 Architecture.

3.3 Generazione delle Immagini

Una volta ottenuto il testo sintetizzato, questo viene passato a Stable Diffusion, un modello di intelligenza artificiale per la generazione di immagini creato da CompVis, Stability AI e LAION. Stable Diffusion utilizza una tecnica di generazione di immagini chiamata diffusione latente, come descritto nel documento "High-Resolution Image Synthesis with Latent Diffusion Models."[6] La sua architettura ad encoder-decoder comprende un decoder UNet e un encoder di testo CLIP, ed è stata addestrata su un vasto dataset LAION-5B per generare immagini 512×512 che corrispondono ai prompt testuali, ottimizzando l'utilizzo delle GPU consumer.

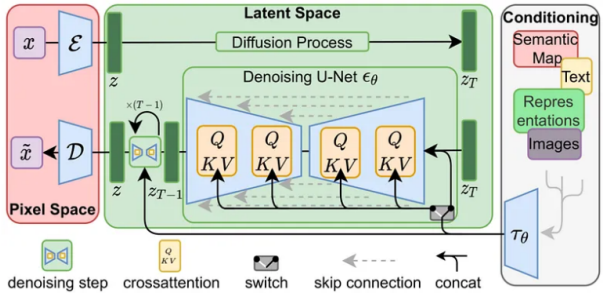

Figure 3: Stable Diffusion Architecture.

Il nostro approccio sfrutta le innovazioni di Stable Diffusion, come la diffusione latente e CLIP, per creare immagini coerenti a partire dal testo sintetizzato. Tuttavia, per garantire un'implementazione efficace e tempestiva, sono state apportate diverse modifiche alla pipeline di generazione delle immagini:

- **Riduzione dei Passi di Inferenza**: Il numero di passi di inferenza è stato limitato a 25, ottenendo così una significativa velocizzazione nei tempi di elaborazione. Questa ottimizzazione è cruciale per consentire una generazione di immagini più rapida e fluida durante le performance dal vivo.
- **Utilizzo della GPU**: La pipeline è stata configurata per sfruttare appieno la potenza della GPU, accelerando i calcoli rispetto all'uso della CPU. Questo ha permesso di gestire in modo più efficiente il carico computazionale, rendendo possibile la generazione di immagini in tempo reale.
- **Disabilitazione del Safety Checker**: La disattivazione del controllo di sicurezza si è rivelata necessaria, poiché l'attivazione di questa funzione generava frequentemente immagini nere a causa di parole sensibili nei testi.
- **Sequential CPU Offload**: Questa funzionalità è stata attivata per migliorare la gestione della memoria e ottimizzare l'inferenza. Grazie a questa configurazione, l'intero processo risulta più fluido e adatto per una generazione continua e in tempo reale delle immagini.

Attraverso queste ottimizzazioni, il nostro sistema è in grado di generare immagini in modo efficace e coerente con il testo, contribuendo a migliorare l'esperienza visiva durante le esecuzioni dal vivo.

4 Risultati

Qui di seguito si analizzeranno i risultati ottenuti da ogni step della pipeline definita nella sezione precedente. Tuttavia, essendo la valutazione delle immagini un processo intrinsecamente soggettivo, non è semplice affermare con certezza se il sistema genera immagini che possano ritenersi coerenti con quanto si sta ascoltando.

Trascrizione del Testo: Analizzando gli output del modello Whisper, si può osservare che il più delle volte il modello ha dimostrato di funzionare bene nel riconoscimento vocale automatico restituendo correttamente la trascrizione del testo. In alcune casistiche, però, ha incontrato alcune difficoltà dovute alla natura delle canzoni. Poiché le tracce musicali includono un sottofondo sonoro e frequenti variazioni nel tono della voce, è possibile che alcune parole vengano trascritte in modo errato o omesse. Questa problematica può influenzare la qualità del testo di input per le fasi successive della pipeline. Prendendo in esame i primi 20 secondi del brano **Shine Like a Diamond** di Rihanna, il testo trascritto è risultato essere: "*Shame right like a diamond If I lie in the beautiful sea I just will be happy You and I, you and I.*" Rispetto al testo originale: "*Shine bright like a diamond, Find light in the beautiful sea, I choose to be happy, You and I, you and I.*" Si notano alcune discrepanze nelle parole trascritte, dovute probabilmente al sottofondo musicale e alla pronuncia delle parole, che possono generare interpretazioni errate o imprecisioni. Nel caso del brano **Yellow Submarine** dei Beatles, Whisper ha prodotto il testo seguente: "*In the town where I was born, lived the man who sailed to sea, and*

he told us of his life, in the land of submarines." Rispetto al testo originale: "*In the town where I was born, Lived a man who sailed to sea, And he told us of his life, In the land of submarines.*" Il risultato è molto vicino all'originale, con solo lievi variazioni, dimostrando una maggiore precisione per alcuni brani rispetto ad altri. Questo suggerisce che la qualità della trascrizione può variare in base a fattori come la chiarezza della pronuncia e la complessità del sottofondo musicale.

Sintesi del Testo: Il secondo step, che implica la sintesi del testo, presenta delle complessità. Utilizzando il codice per estrarre segmenti audio di 20 secondi, abbiamo accumulato il testo trascritto per generare un'immagine ad intervalli regolari. Tuttavia, si è notato che, a un certo punto, il testo sintetizzato risultava ripetitivo e invariato. Questo comportamento è preoccupante, poiché comporta che l'input per il modello Stable Diffusion rimanga costante, compromettendo la varietà delle immagini generate. Le cause di questa ripetizione potrebbero risiedere nell'accumulo del testo o nel modo in cui il modello T5 sintetizza il contenuto. Infatti il modello T5 non prevede alcun modo per randomizzare la sintesi che produce, per ovviare a questo problema si potrebbe usare qualche altro modello che implementi tale modalità.

Generazione delle Immagini: Per quanto riguarda la generazione delle immagini, il modello Stable Diffusion ha prodotto risultati visivamente suggestivi. Le immagini create riflettono spesso i temi e le emozioni evocate dai testi, ma, come è prevedibile, presentano sempre delle anomalie tipiche dei modelli generativi. L'analisi ha evidenziato che, mentre l'uso di un numero ridotto di passi di inferenza (25) ha accelerato i tempi di generazione, questa scelta ha anche comportato una diminuzione della qualità visiva delle immagini. In uno scenario di utilizzo in tempo reale, la priorità è stata data alla velocità, ma in situazioni in cui il tempo non è un fattore critico, l'aumento dei passi di inferenza potrebbe migliorare significativamente l'estetica delle immagini generate.
Infine, a titolo di esempio, si presenteranno alcune istantanee delle immagini prodotte per il brano **Yellow Submarine** dei Beatles.

Figure 4: Esempio di immagini generate.

5 Conclusioni

Limitazioni

Come dimostrano i risultati, una delle principali limitazioni di questo approccio è la necessità di ridurre il numero di passi di inferenza per mantenere un tempo di generazione delle immagini accettabile. Sebbene questa riduzione sia necessaria per ottimizzare le prestazioni, si traduce in immagini di qualità inferiore rispetto a quelle che si potrebbero ottenere con un hardware più performante. La configurazione attuale, che comprende una scheda grafica e un processore non recenti, limita di fatto la possibilità di utilizzare modelli più complessi e più fasi di inferenza senza compromettere la velocità di elaborazione. Un'altra limitazione è rappresentata dal fatto che la generazione dell'immagine si basa esclusivamente sul testo, senza prendere in considerazione gli elementi musicali, come la melodia e l'intonazione, che contribuiscono a definire il contesto emotivo della scena. In effetti, l'assenza della componente musicale può portare alla creazione di immagini che non riflettono appieno il tono emotivo della canzone, poiché il testo da solo non è sempre sufficiente a trasmettere emozioni come tristezza o gioia, che spesso derivano dal sottofondo musicale e dall'interpretazione vocale dell'artista.

Sviluppi futuri

Data l'analisi delle limitazioni che si sono riscontrate, sicuramente una futura implementazione della pipeline deve considerare la parte musicale del brano, andando ad utilizzare dei modelli che consentono di analizzare gli elementi musicali, come la melodia e l'intonazione dell'artista, e di estrarne caratteristiche significative da passare ad un modello di generazione delle immagini che non prenda come input il solo testo, ma anche queste caratteristiche significative, permettendo di generare immagini più coerenti e fedeli all'atmosfera del brano, consentendo di visualizzare non solo il contenuto verable del brano ma anche le emozioni

suscitate dall'accompagnamento musicale.

References

[1] Yue Qiu and Hirokatsu Kataoka. "Image Generation Associated With Music Data". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2018, pp. 2510–2513. DOI: 10.1109/CVPRW.2018.00032.

[2] Jiaodong Tan and Mathis Antony. "Music Generation from Visual Data". In: *Proceedings of the 20th International Conference on Computational Creativity (ICCC)*. 2024, pp. 12–20. DOI: 10.48550/arXiv.2407.05584.

[3] Meng Yang, Maria Teresa Llano, and Jon McCormack. "Exploring Real-Time Music-to-Image Systems for Creative Inspiration in Music Creation". In: *15th International Conference on Computational Creativity (ICCC)*. 2024, pp. 1–8. DOI: 10.48550/arXiv.2407.05584.

[4] Alec Radford et al. *Whisper: A General-Purpose Speech Recognition Model*. Accessed: 2024-11-01. 2022. URL: https://openai.com/research/whisper.

[5] Colin Raffel, Chris Shinn, Roberts, et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Proceedings of the 37th International Conference on Machine Learning*. Accessed: 2024-11-01. PMLR, 2020. URL: http://proceedings.mlr.press/v119/raffel20a.html.

[6] Felix Rombach, Blattmann, et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Accessed: 2024-11-01. 2022, pp. 10823–10833. URL: https://arxiv.org/abs/2112.10752.