Background Music Generation using Movies Scripts

Nicola Pio Santorsa
Università degli Studi di Salerno
Curriculum Data Science and Machine Learning
Salerno, Italy
n.santorsa@studenti.unisa.it

Grazia Margarella
Università degli Studi di Salerno
Curriculum Data Science and Machine Learning
Salerno, Italy
g.margarella2@studenti.unisa.it

***Abstract*—Nell'ambito di questo report è descritto il processo di prototipazione di una pipeline per risolvere il task di background music generation per scene cinematografiche. Per tale scopo, è stato utilizzato un dataset costituito da sceneggiature di diversi film, suddivisi per scena, con informazioni riguardanti la messa in scena e i dialoghi. Questi dati sono stati analizzati ed elaborati tramite due LLM. Il primo modello si occupa di analizzare il testo della sceneggiatura e di estrarre informazioni rilevanti per la generazione di un brano da porre in background, come un genere musicale e caratteristiche descrittive dell'atmosfera e del contesto della scena. Successivamente un secondo modello utilizza le informazioni estratte per generare una melodia da poter utilizzare come sottofondo musicale alla scena designata.**

***Index Terms*—Large Language Models, Music Background Generation, Film Scripts**

I. Introduzione

La storia del cinema e del mondo audiovisivo hanno avuto come uno dei suoi linguaggi fondamentali la musica. Essa infatti il più delle volte è la vera protagonista di alcune scene che ancor oggi ricordiamo dei nostri film preferiti, dando carattere e significato a diversi tipi di immagini. D'altra parte, anche le melodie utilizzate in sottofondo in video disponibili sulla piattaforma Youtube, nonostante possano sembrare meno nobili rispetto alle iconiche colonne sonore che vengono in mente ad ognuno di noi, esse rappresentano uno strumento essenziale per enfatizzare concetti, supportare la narrazione, catturare l'attenzione dello spettatore o dare altre chiavi di lettura. Con il crescere del mondo audiovisivo con prodotti come film, serie TV, documentari, scene cinematiche di videogiochi o video pubblicati su diverse piattaforme o social, la richiesta di musica senza diritti d'autore è sempre più in crescita.

Per rispondere a questa richiesta si utilizzano spesso modelli generativi capaci di comporre melodie inerenti con le descrizioni date in input dagli utenti. In questo contesto si colloca questo report, in cui si va a descrivere il processo di sviluppo di un'applicazione che, a partire dalle sceneggiature cinematografiche, genera melodie da porre in sottofondo a scene di film o serie tv. Per far ciò si è sviluppata una pipeline caratterizzata da due Large Language Models, Gemma 2 e MusicGen, uno per analizzare il testo della sceneggiatura e l'altro per generare la melodia. Il codice del seguente progetto è disponibile Open al seguente link.
La struttura del presente report viene riportata qui di seguito:

- Nella Sezione II, si analizza l'attuale stato dell'arte in ambito Background Music Generation, approfondendo alcuni approcci video-based e alcuni modelli generativi che permettono di analizzare i video. Si vedranno inoltre anche alcuni studi che si sono concentrati sullo studio della sceneggiatura con lo scopo di generare melodie da essa;
- Nella Sezione III, si analizza il metodo proposto per generare le melodie da porre come sottofondo in scene cinematografiche, descrivendo i dataset e i modelli utilizzati per tale scopo;
- Nella Sezione IV, si analizzano i risultati conseguiti, verificando la coerenza delle caratteristiche estratte dalla scena e l'attinenza della melodia generata con la stessa;
- Infine, nella Sezione V, si discute il processo traendo le dovute conclusioni, mostrando le limitazioni di questo prototipo e si delineano gli sviluppi futuri.

II. Stato dell'arte

A. Music Background Generation from Videos

Molti degli studi presenti in letteratura in ambito *Music Background Generation* si concentrano sull'analisi di input di tipo multimediale, nello specifico di input video. In Shangzhe Di et al. [1] si estraggono a partire da un video le relazioni ritmiche presenti tra la parte visuale del video e la musica di sottofondo. In questa analisi si vanno a correlare alcune proprietà del video come il timing, la velocità e le caratteristiche salienti del movimento, ad attributi della melodia come il ritmo, la densità e l'intensità delle note. Per estrarre queste proprietà dai video, per poi generare una melodia, hanno definito un modello basato sull'architettura Transformers che successivamente con altre informazioni date in input dall'utente, come il genere musicale e i tipi di strumenti musicali utilizzati, permettono la generazione di una melodia compatibile con il video analizzato. Kun Su et al. [2] hanno sviluppato un modello autoregressivo che prende in considerazione le forme d'onda della melodia di un video e le correla a diverse feature visuali estratte da frame del video in input, utilizzando un modello LLM interrogato tramite prompt testuale. Jaeyong Kang et al. [3] hanno sviluppato un sistema per estrarre diverse caratteristiche a partire dai video come la semantica, scene offset, movimento, e le emozioni, ed hanno utilizzato queste per la generazione della melodia tramite il loro modello generativo basato su un'architettura Transformer Multimodale.

	Scene_Names	Scene_action	Scene_Characters	Scene_Dialogue	Contents
0	EXT. THRONE ROOM, SPACE NIGHT	Kneeling behind a THRONE, a CLOTHED, ARMORED F...	None	None	Kneeling behind a THRONE, a CLOTHED, ARMORED ...
1	EXT. S.H.I.E.L.D. PROJECT P.E.G.A.S.U.S FACILI...	Out in the NEW MEXICO desert, a remote researc...	None	None	Out in the NEW MEXICO desert, a remote resear...
2	EXT. HELICOPTER PAD CONTINUOUS	Standing a few yards from the landing pad, A F...	[NICK FURY, AGENT PHIL COULSON]	[How bad is it?, That's the problem, sir. We d...	Standing a few yards from the landing pad, A ...
3	INT. FACILITY FLOOR NIGHT	Agent Coulson leads Hill and Fury through the ...	[AGENT PHIL COULSON, NICK FURY, AGENT PHIL COU...	[Dr. Selvig read an energy surge from the Tess...	Agent Coulson leads Hill and Fury through the...
4	INT. NASA SPACE RADIATION FACILITY, VACUUM CHA...	Fury enters the lab facility where the Tessera...	[NICK FURY, SELVIG, NICK FURY, SELVIG, NICK FU...	[Talk to me, doctor. DR. ERIK SELVIG emerges f...	Fury enters the lab facility where the Tesser...
...	...	...	...	...	...
211	INT. SHIELD ANALYTICAL ROOM DAY	In TV news montage about THE AVENGERS, we see ...	[SENATOR BOYNTON, WAITRESS]	[These so called heroes have to be held respon...	In TV news montage about THE AVENGERS, we see...
212	EXT. CENTRAL PARK DAY	The Avengers take Thor and Loki, who is handcu...	None	None	The Avengers take Thor and Loki, who is handc...
213	INT. SHIELD ANALYTICAL ROOM DAY	Fury is facing once more members of the WORLD ...	[NICK FURY, NICK FURY, NICK FURY, NICK FURY, N...	[I am not currently tracking their whereabouts...	Fury is facing once more members of the WORLD...
214	INT. HELICARRIER BRIDGE DAY	Fury and Agent Hill walk together, toward the ...	[AGENT MARIA HILL, NICK FURY, AGENT MARIA HILL...	[Sir, how does it work now? They have gone the...	Fury and Agent Hill walk together, toward the...
215	INT. STARK PENTHOUSE DAY	TONY AND PEPPER UNVEIL A NEW DESIGN FOR STARK ...	None	None	TONY AND PEPPER UNVEIL A NEW DESIGN FOR STARK...

Fig. 1. Esempio dataset film "The Avengers (2012)"

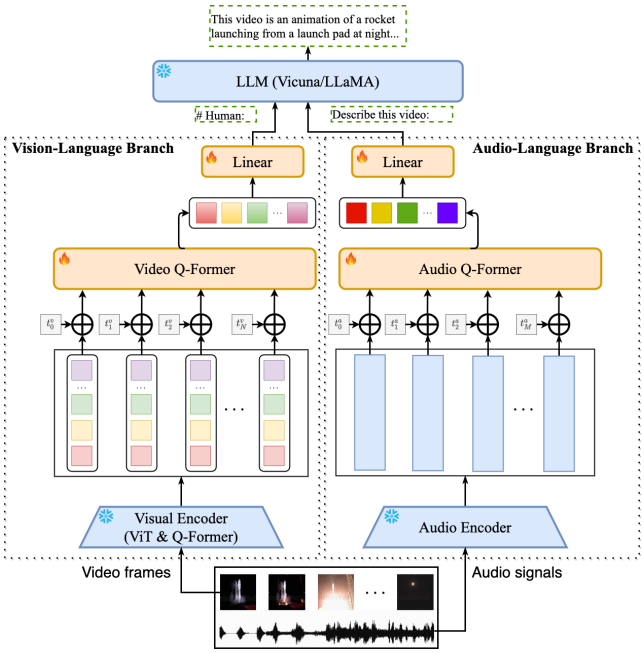

Fig. 2. Architettura VideoLLAMA

1) Video Analyzer LLM: Nell'attuale stato dell'arte, sono presenti diversi modelli che permettono l'estrazione delle caratteristiche visuali a partire da un input video, ma molti di questi non sono disponibili Open Source, come per esempio il modello **GPT-4-Vision** sviluppato da *OpenAI*. Tra i modelli Open Source appaiono due LLM sviluppati dal *Language Technology Lab at Alibaba DAMO Academy*, che sono rispettivamente VideoLLAMA [4] e la sua versione successiva, VideoLLAMA2 [5]. Come viene mostrato nelle figure 2 e 3, le architetture dei due modelli hanno il compito di analizzare separatamente la parte visuale e la parte audio di un input video. Dopo questa analisi i modelli possono essere interrogati, tramite prompt testuale, sugli eventi descritti dal video, in modo da poter estrarre determinate informazioni. Entrambi i modelli dispongono di una demo gratuita tramite il sito Hugging Face per testare le loro funzionalità, disponibile al link la Demo del modello VideoLLAMA2.

B. Analisi degli script cinematografici

Altri studi si sono concentrati invece sull'analisi testuale per la generazione di melodie per scene di film. Alexis Kirke e E. Miranda [6] hanno sviluppato un sistema che supporta i compositori durante la generazione di colonne sonore per film. Il sistema effettua un analisi automatica del testo che va ad approssimare la struttura dello script e tramite l'analisi delle parole estrae le emozioni che si vogliono trasmettere. Queste caratteristiche estratte vengono poi utilizzate come input ad un algoritmo creativo che suggerisce al compositore degli Sketch musicali per le parti della sceneggiatura. Vishruth Veerendranath et al. [7] hanno definito una pipeline che, tramite le sceneggiature dei film, estrae in una prima fase i sentimenti che vengono mappati in uno spazio continuo del tipo *Valence-Arousal*. Nella seconda fase si utilizzano i vettori generati al passo precedente come input ad un modello generativo condizionale per generare una melodia tramite pianoforte in formato MIDI.

III. Metodologia

A. Dataset

I dataset che si sono utilizzati durante questo studio provengono dal sito IMSDB, il quale è uno dei database più grandi disponibili online di script cinematografici. Per estrarre i copioni dei film si è utilizzato il codice disponibile al seguente link, il quale implementa un web scraper per estrarre le sceneggiature cinematografiche da IMSDB, le quali sono state successivamente segmentate in scene utilizzando un segmenter personalizzato. I file così generati sono stati organizzati in oggetti di tipo DataFrame, utilizzando la libreria *pandas* di Python. Ogni dataframe così generato è strutturato da n righe quante sono le scene del film, e 5 colonne in cui vengono descritte caratteristiche della scena, e sono:

- **Scene_Names**: Una nomenclatura utilizzata all'interno del copione per identificare univocamente la scena. In essa sono inoltre presenti dettagli sull'ambientazione della scena;
- **Scene_Action**: Una breve descrizione sullo svolgimento della scena, inoltre include dettagli sull'ambiente mostrato nella scena, i personaggi presenti in essa ed azioni che stanno svolgendo;

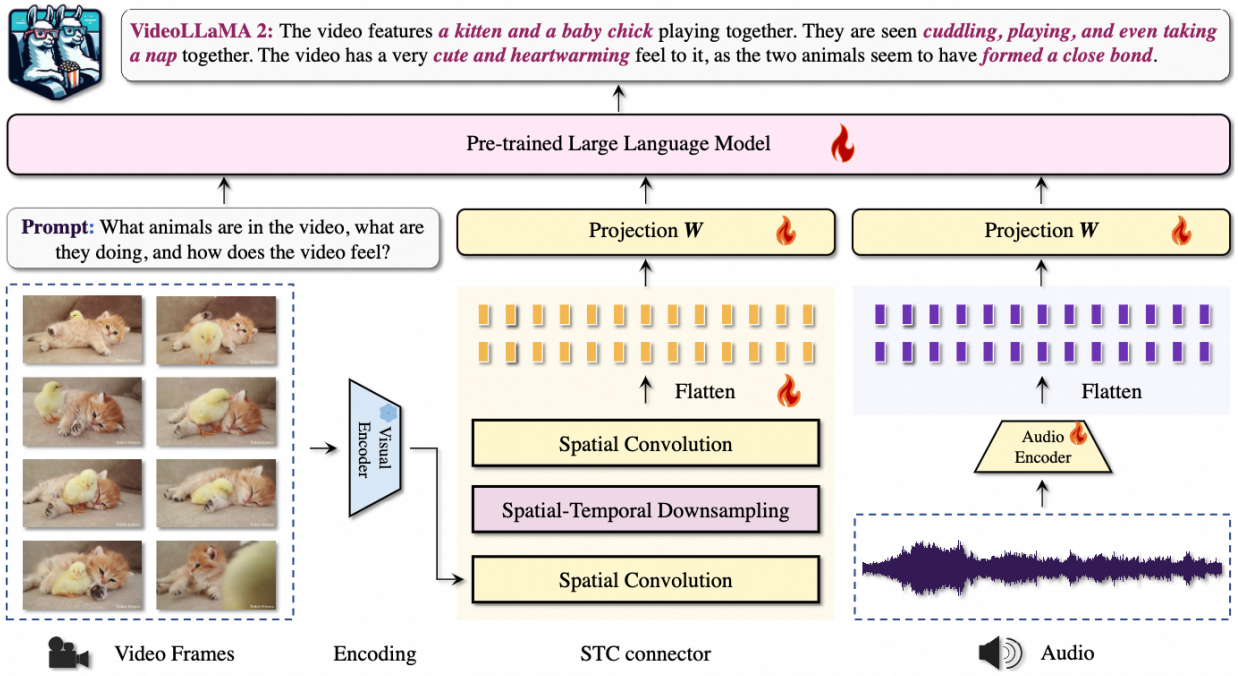

Fig. 3. Architettura VideoLLAMA2

- **Scene_Characters**: Lista di personaggi presenti nella scena descritta;
- **Scene_Dialogue**: Dialoghi tra personaggi nella scena descritta;
- **Contents**: Un testo generato dall'unione del campo *Scene_Action* e *Scene_Dialogue*;

Viene mostrata in figura 1 un esempio sulla struttura del dataset. I film che abbiamo selezionato per questa analisi rientrano quasi tutti nelle categorie *Action e Fantascienza*, dal momento che sono i generi più rappresentati nel dataset e più sfidanti per il task generativo. Tra questi sono presenti film come *Star Wars, Avengers, Toy Story*, etc., ma la scelta di nuovi film dal dataset è facilmente implementabile.

B. Pipeline

La pipeline che è stata sviluppata si suddivide in due step fondamentali: nel primo si analizza la descrizione testuale della scena per estrarre caratteristiche da poi utilizzare per generare la melodia rappresentativa nel secondo step. Un immagine sull'architettura della pipeline sviluppata è mostrata in figura 5

1) Analisi degli Script: Durante questa prima fase il task da compiere è quello di estrarre caratteristiche dalla scena che possano essere poi utili per generare una melodia adatta alle azioni che si stanno svolgendo nel film nella scena selezionata. Per poter estrarre queste caratteristiche necessitiamo di una descrizione fedele di tutto ciò che appare nella scena cinematografica a partire dall'ambiente in cui si svolge, passando poi alle azioni che stanno effettuando i personaggi, e terminando con i dialoghi. Tutte queste informazioni elencate sono contenute all'interno del dataset tramite la feature *Content*. Per effettuare l'analisi testuale di questa descrizione si è utilizzato un modello **Gemma 2** [8], il quale è stato rilasciato a Febbraio 2024 da *Google*, ed è un modello light-weighting caratterizzato da circa 2.5 bilioni di parametri[1]. Per estrarre le caratteristiche della scena si è dato come input al modello gemma un prompt così composto:

"Imagine you have to write a song about the scene described, describe with 3 words the song which will fit this scene, the musical genre and the main emotion. + *Contents*".

Le caratteristiche fondamentali per il passo successivo estratte tramite questo metodo sono principalmente il genere musicale consigliato per la scena e le emozioni che essa suscita. Da notare come le emozioni indicate non siano una semplice classificazione, ma una descrizione emotiva della scena.

2) Generazione della Melodia: Dati gli attributi risultati dal passo precedente, in questa fase si procederà a generare la melodia condizionata dalle informazioni in input. Il modello utilizzato per generare la melodia è **MusicGen** [9], sviluppato da *Meta* e rilasciato nel 2023. Per il task di generazione, il modello richiede come input un prompt testuale in cui si descriva la melodia che si deve produrre. Il prompt che si è dato in input al modello è composto dal genere musicale consigliato e dalle emozioni suscitate dalla scena. Sono diversi i parametri configurativi del modello che gestiscono la generazione della melodia, tra i più importanti è presente il

[1]Il modello è disponibile OpenSource tramite il sito Hugging Face al seguente link

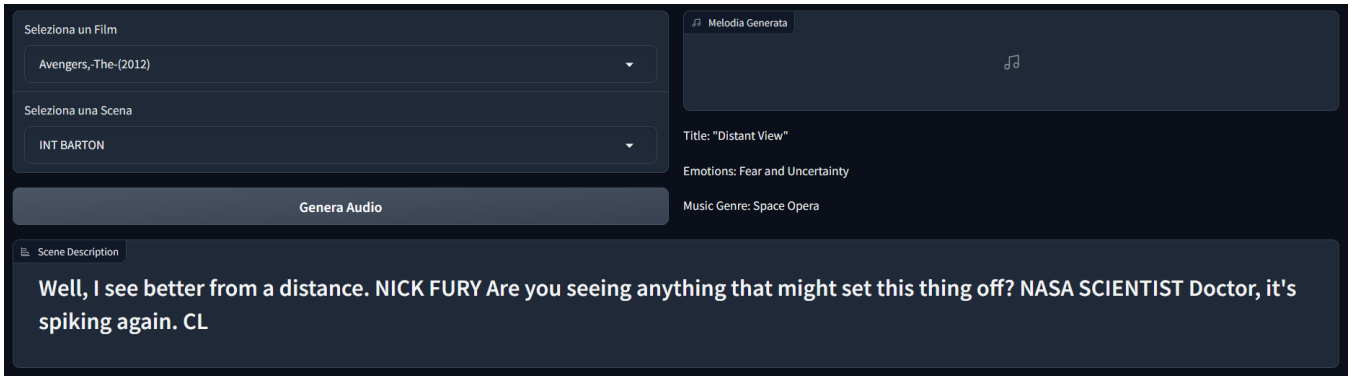

Fig. 4. Interfaccia implementata utilizzando Gradio

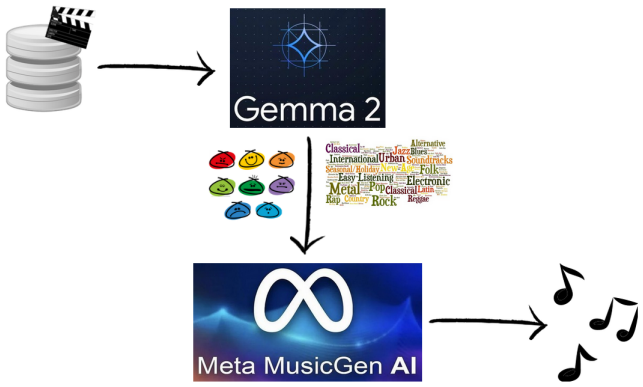

Fig. 5. Architettura della Pipeline

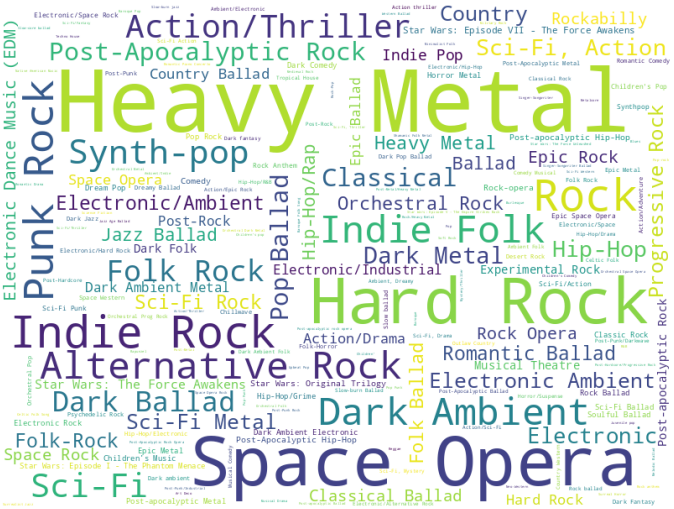

Fig. 6. Word Cloud Generi Musicali

parametro *max_new_tokens*, il quale gestisce la durata della melodia generata. L'approccio che si è scelto di adottare nel seguente studio è stato quello di impostare il valore del seguente parametro a 1500, che corrisponde ad una durata della melodia di 30 secondi.

C. Interfaccia Grafica

Per facilitare l'interazione dell'utente con la pipeline sviluppata si è creata un'interfaccia grafica utilizzando il framework **Gradio**, tramite la quale si può selezionare il film e la scena che si vogliono analizzare. Dopo aver generato la melodia l'interfaccia consente di riprodurla senza alcun bisogno di doverla scaricare, inoltre vengono mostrate sull'interfaccia le caratteristiche estratte, come il genere musicale consigliato, le emozioni che la scena suscita, un eventuale titolo per la melodia, e la stringa testuale analizzata dalla pipeline. L'interfaccia cosi' composta viene mostrata in figura 4.

IV. Risultati

Qui di seguito si analizzeranno i risultati ottenuti da ogni step della pipeline che si è definita nella sezione precedente. Essendo l'ambito della Music Background Generation ancora agli albori, non esistono delle metriche standardizzate per valutare la qualità dei risultati prodotti, l'unica metrica che è stata introdotta è la **Video-Music CLIP Precision(VMCP)** [10] che però utilizza una valutazione basata sui video che non viene utilizzata nel seguente studio. Per i motivi descritti quindi le valutazioni che saranno date in questa sezione saranno strettamente basate su valutazioni soggettive. In futuro nel caso si volesse approfondire la qualità delle melodie generate si potrebbero validare i risultati tramite uno user study.

A. Estrazione delle Caratteristiche

Se si analizzano gli output del modello Gemma, si può osservare che il più delle volte il modello restituisca una struttura precisa dell'output, composta dal genere musicale suggerito e dalle emozioni. In alcune casistiche il modello genera anche un titolo per la melodia, ma non essendo una caratteristica richiesta dal prompt, questo non è sempre generato. Infine si può osservare che in pochi casi, il modello non restituisce nessuna delle caratteristiche richieste, probabilmente causato dall'elevata complessità o lunghezza del testo che descrive la scena. Analizzando i generi suggeriti dal modello, come si può osservare tramite la word cloud in Figura 6, la prevalenza di film di tipo Action ha portato ad un suggerimento maggiore di generi di tipo Rock e Metal che risultano più adeguati in

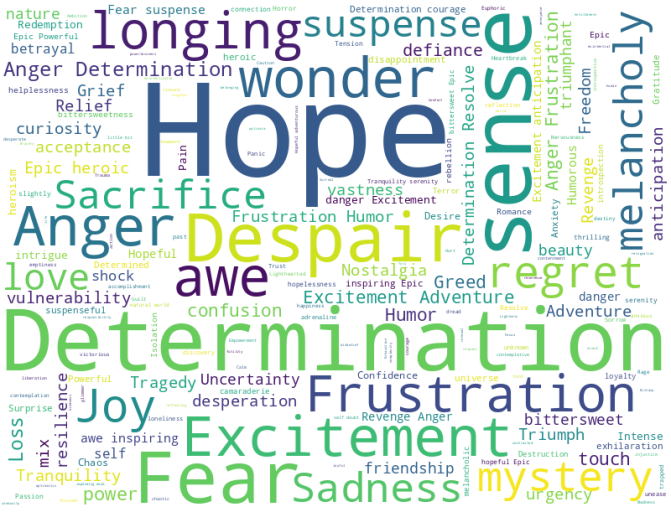

Fig. 7. Word Cloud Emozioni

scene movimentate in cui si vuole trasmettere all'utente una forte emozione e mantenere alta la sua attenzione, e inoltre un'alta presenza di genere *Space Opera* dovuto dal numero elevato di scene appartenenti alla saga *Star Wars*. Analizzando le emozioni estratte dal modello, come si può osservare tramite la word cloud in Figura 7, si può notare una grande diversità tra le emozioni rilevate dal modello, tra le quali molte che esprimono concetti molti più complessi della semplice paura o gioia, il che ci mostra il grande livello di approfondimento del modello di emozioni complesse e di difficile rilevazione. Inoltre per queste si è potuto anche rilevare una spiccata capacità del modello nel descrivere con grande accuratezza particolari scenari, come quelli presenti nella sceneggiatura del famoso film *Titanic*, in cui il modello per descrivere le emozioni trasmesse da scene in cui viene mostrato il mare utilizza la seguente descrizione testuale: *A sense of awe and wonder at the vastness and mystery of the deep sea*. In generale, non considerando le scene in cui il modello non è riuscito ad estrarre gli attributi richiesti, esso riesce con successo a catturare molti degli aspetti complessi della scena e riesce a descriverla con efficacia utilizzando poche parole per descrivere i sentimenti che essa suscita. Una problematica che si è rilevata nella generazione delle emozioni è basata sulla parziale inesattezza di alcune emozioni che la scena suscita, questa è dovuta dall'analisi isolata della scena, ovvero dalla non conoscenza da parte del modello del contesto delle scene precedenti che possono influenzare parzialmente le emozioni suscitate dalla scena in questione.

B. Melodie Generate

Analizzando le melodie generate dal modello MusicGen, si è subito potuto osservare la capacita' del modello di interpretare al meglio le istruzioni date tramite prompt, rispettando in primis il genere musicale richiesto, ma sopratutto riuscendo ad esprimere anche emozioni complesse con melodie di breve durata. Se si prendono in considerazione scene di suspense dei film, si è potuto osservare l'inserimento di un sottofondo nella melodia tipico dei film in cui si cerca di incutere ansia e si tenta di tenere l'attenzione dello spettatore alta. Nonostante gli sviluppatori del modello abbiano identificato come una delle problematiche la poca diversità di tracce utilizzate per l'addestramento del modello, le quali erano molto più sbilanciate verso il genere western, nelle melodie che si sono generate durante questo progetto non si è rilevato alcun problema di questo tipo, anzi, come detto in precedenza, il modello è riuscito a generare melodie che rispettavano anche emozioni più complesse che difficilmente sono rappresentabili.

V. Conclusioni

A. Limitazioni

Da come si evince dai risultati che si sono analizzati, la limitazione più grande di un approccio di questo tipo è la sola analisi del testo e non anche della parte visiva della scena del film. Purtroppo con le risorse hardware disponibili non si è potuto integrare anche uno step di analisi video, utilizzando i modelli VideoLLAMA, poiché essi utilizzano un numero elevato di parametri, circa 7 Bilioni, che corrispondono circa a 28GB di RAM per il solo mantenimento del modello in memoria, da aggiungere in seconda battuta le risorse necessarie per l'inferenza del modello. Oltre questa limitazione, un'ulteriore problematica che si è riscontrata si basa sulla non conoscenza del modello sui fatti accaduti nelle scene precedenti, le quali giocano un ruolo fondamentale nella definizione dei sentimenti espressi da una scena. Si pensi ad un flashback infantile il cui sentimento principale espresso è la gioia. Se questo venisse "isolato" in un unica scena, il modello identificherebbe questa come una scena gioiosa, se però la scena che l'ha preceduta fosse una scena colma di tristezza in cui il personaggio ripensa ai momenti felici della sua infanzia, l'influenza della scena precedente andrebbe a stravolgere completamente il quadro emotivo della scena successiva portando anche un senso di malinconia.

B. Sviluppi Futuri

Data l'analisi delle limitazioni che si sono riscontrate, sicuramente una futura re-implementazione della pipeline deve considerare la parte multimediale della scena, andando quindi ad utilizzare modelli generativi come VideoLLAMA che consentono di analizzare video e di estrarne caratteristiche utilizzando lo stesso approccio adottato con Gemma. L'aggiunta di questo step nella pipeline non sarebbe per forza come sostituzione dell'analisi testuale della sceneggiatura, anzi esso si potrebbe utilizzare a supporto, in modo tale da estrarre caratteristiche sia dal video che dal testo, cosi' effettuando un'unione di queste informazioni estratte e successivamente generare tramite MusicGen una melodia che riesca ad abbracciare sia le caratteristiche video che testuali della scena. Inoltre si potrebbe pensare anche all'introduzione di una memoria dei modelli che estraggono le caratteristiche, in modo che essi conservino le informazioni delle scene precedenti e tramite le informazioni passate e le informazioni presenti riescano ad estrarre con maggiore efficienza le emozioni che si trasmettono a partire dalla scena cinematografica. L'idea per l'utilizzo

di questa pipeline è sicuramente di supporto a personale esperto del montaggio video per l'introduzione di musiche di background in base al contesto, e ciò potrebbe essere realizzato oltre tramite la piattaforma web accennata nella definizione di questo progetto, ma anche tramite plugin a software di editing video.

REFERENCES

[1] S. Di, Z. Jiang, S. Liu, Z. Wang, L. Zhu, Z. He, H. Liu, and S. Yan, "Video background music generation with controllable music transformer," in *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1145/3474085.3475195

[2] K. Su, J. Y. Li, Q. Huang, D. Kuzmin, J. Lee, C. Donahue, F. Sha, A. Jansen, Y. Wang, M. Verzetti *et al.*, "V2meow: Meowing to the visual beat via video-to-music generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4952–4960.

[3] J. Kang, S. Poria, and D. Herremans, "Video2music: Suitable music generation from videos using an affective multimodal transformer model," *Expert Systems with Applications*, vol. 249, p. 123640, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424005062

[4] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," 2023. [Online]. Available: https://arxiv.org/abs/2306.02858

[5] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, and L. Bing, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," 2024. [Online]. Available: https://arxiv.org/abs/2406.07476

[6] A. Kirke and E. Miranda, "Aiding soundtrack composer creativity through automated film script-profiled algorithmic composition," *Journal of Creative Music Systems*, vol. 1, no. 2, 2017.

[7] V. Veerendranath, V. Masti, U. Gupta, H. Chaudhuri, and G. Srinivasa, "Scriptones: Sentiment-conditioned music generation for movie scripts," in *Proceedings of the Third International Conference on AI-ML Systems*, 2023, pp. 1–6.

[8] Gemma: Introducing new state-of-the-art open models. [Online]. Available: https://blog.google/technology/developers/gemma-open-models/

[9] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[10] L. Zhuo, Z. Wang, B. Wang, Y. Liao, C. Bao, S. Peng, S. Han, A. Zhang, F. Fang, and S. Liu, "Video background music generation: Dataset, method and evaluation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 15 637–15 647.