Universidade do Minho
Escola de Engenharia

Geração de Música com Modelos de Linguagem

Aprendizagem Profunda
Mestrado em Engenharia Informática

Grupo 12

Gabriela Santos Ferreira da Cunha - pg53829

Millena de Freitas Santos - pg54107

Nuno Guilherme Cruz Varela - pg54117



maio, 2024

Conteúdo

1 Introdução

No âmbito da unidade curricular de Aprendizagem Profunda foi-nos proposta a conceção e desenvolvimento de um projeto de *deep learning* utilizando as técnicas abordadas ao longo do semestre. Desta forma, o projeto desenvolvido consiste na exploração e refinamento de um modelo de linguagem para geração de música, visando a especialização na geração de músicas de um género específico. Neste sentido, o modelo utilizado foi o MusicGen, desenvolvido pela equipa de investigação da MetaAI, e o género escolhido foram as desgarradas.

2 Contextualização e Motivação

2.1 Exploração do Modelo

Para realizar a especialização do modelo, é essencial primeiro examinar os conjuntos de dados de músicas disponíveis, bem como o estado atual do próprio modelo. É necessário identificar áreas em que o modelo esteja a produzir resultados abaixo do esperado e determinar se há dados suficientes disponíveis para aprimorar o seu desempenho nesse estilo específico. Esta primeira etapa é fundamental, sendo que só a partir desta exploração e avaliação inicial é que conseguimos direcionar efetivamente o projeto e definir objetivos mais concretos para a otimização, isto é, definindo o género musical sobre o qual pretendemos melhorar a capacidade do modelo em gerar músicas com qualidade e autenticidade.

Inicialmente, começamos por explorar géneros musicais mais comuns, como o caso do *pop* e da música clássica, onde concluímos que os resultados obtidos pelo modelo eram satisfatórios e, portanto, conduzir a especialização para esses estilos não traduziria uma grande adição ao modelo original. Neste sentido, vimos a necessidade de ampliar os horizontes e considerar a diversidade musical, direcionando a nossa atenção para estilos menos convencionais e mais específicos, que não necessariamente ressoassem com todas as culturas de maneira universal.

As desgarradas consistem num canto improvisado típico da música tradicional portuguesa que é feito em duelo entre 2 ou mais cantores. Este tipo de músicas é geralmente tocado com instrumentos tradicionais, como a concertina ou o acordeão. Ao testarmos *prompts* como "desgarrada", "desgarrada concertina" ou "desgarrada accordion", percebemos que os resultados do modelo eram pouco relacionados com este género musical e que esta seria a direção indicada para o refinamento do modelo.

2.2 Objetivos

- **Especialização do MusicGen:** Garantir a especialização do modelo MusicGen na geração de músicas de desgarrada, através do *fine-tuning* e treino com conjuntos de dados específicos deste género musical;

- **Avaliação das músicas geradas:** Desenvolver um sistema abrangente para a avaliação das músicas geradas, permitindo a recolha de *feedback* detalhado sobre diferentes aspetos das composições, como originalidade, qualidade e fidelidade ao género.

2.3 Abordagem

- **Exploração inicial do modelo:** Avaliação das capacidades do modelo para determinar quais géneros musicais são bem produzidos e quais precisam de ajustes;
- **Escolha do conjunto de dados:** Seleção de músicas do género musical escolhido para treinar o modelo;
- **Tratamento dos dados:** Tratamento dos áudios ao dividi-los em segmentos de 30s, com 32000 Hz e geração de *labels* `.txt`;
- **Treino do modelo:** *Fine-tuning* do modelo com recurso ao *fine-tuner* MusicGen Trainer;
- **Avaliação dos resultados obtidos:** Criação de um formulário por forma a realizar uma avaliação abrangente e imparcial de diferentes aspetos subjetivos;

3 Otimização do Modelo

3.1 Ambiente

A otimização foi realizada no Google Colab, com a GPU T4 disponível na versão gratuita da plataforma. Esta escolha proporcionou acesso a recursos computacionais superiores à disposição do grupo.

No entanto, a versão gratuita do Google Colab impôs algumas limitações significativas, como o tempo de utilização limitado da GPU e desconexões frequentes, o que resultou em interrupções constantes durante o processo de treino. Para contornar essas dificuldades, foi necessário criar novos *notebooks* frequentemente e até utilizar contas Google diferentes.

Para assegurar a persistência e acessibilidade dos dados ao longo do processo, foi utilizado o Google Drive para armazenar os dados de entrada, os segmentos tratados e as suas respetivas *labels*, bem como as músicas geradas.

3.2 Dados

Para recolher os dados necessários para o treino do modelo, utilizamos o Youtube como fonte principal. Foram realizadas pesquisas específicas por vídeos e músicas de desgarradas e extraímos o áudio dos vídeos em formato MP3.

Após o *download* das músicas recolhidas, foi necessário efetuar o tratamento destes ficheiros para cumprir os requisitos sobre o formato dos dados para o treino do MusicGen através do *trainer* utilizado. Primeiramente, os áudios foram segmentados em trechos de 30 segundos utilizando a biblioteca `pydub`. Em seguida, a frequência dos áudios foi ajustada para 32000 Hz.

O nome original dos arquivos de áudio foi extraído para a criação das *labels*. Estas *labels* foram guardadas em ficheiros de texto (`.txt`), associando cada segmento ao seu respetivo nome.

Por fim, para garantir a integridade dos segmentos de áudio, a biblioteca `librosa` foi utilizada para validar a duração e frequência dos mesmos, assegurando que todos cumprissem os requisitos mencionados acima. Este controlo de qualidade foi essencial para manter a consistência dos dados utilizados no treino do modelo.

3.3 Treino

Para a otimização do MusicGen, optámos por utilizar um *trainer* já existente, disponível no GitHub. Esta solução demonstrou-se apropriada principalmente devido às limitações de recursos computacionais disponíveis e à existência de uma solução pré-construída que poderíamos adaptar aos nossos dados e objetivos.

Inicialmente, o repositório foi clonado e os requisitos necessários foram instalados, como por exemplo a biblioteca `wandb`. O estudo da configuração necessária do ambiente foi essencial para garantir que o *trainer* funcionasse corretamente com os dados de entrada.

Para encontrar a melhor combinação de parâmetros de treino, foram testadas diversas configurações de *epochs* e *batch sizes*, variando entre 5 a 50 *epochs* e *batch sizes* de 2 a 4. Após vários testes, a configuração escolhida foi de 50 *epochs* e *batch size* de 2, pois apresentou os melhores resultados em termos de desempenho e qualidade das músicas geradas.

O *trainer* utilizou o modelo pré-treinado MusicGen na versão *small*. Este modelo é mais leve e, portanto, possui menos parâmetros, sendo mais adequado para os recursos computacionais disponíveis. Apesar de ser mais acessível computacionalmente, é importante observar que o modelo *small* não oferece os melhores resultados possíveis em comparação com versões maiores do modelo, de forma a ajustar as nossas expectativas. Após o treino com os dados relativos às desgarradas, o modelo treinado foi guardado para posterior uso e avaliação.

Para a geração das músicas, foram utilizados diversos *prompts*, como "desgarrada", "desgarrada concertina", "desgarrada accordion" e outras combinações. Esta abordagem permitiu a criação de uma série de músicas distintas, onde cada uma foi cuidadosamente ouvida para garantir que não se tratavam de cópias dos dados originais, mas sim criações novas e únicas. As músicas foram,

então, armazenadas no Google Drive com a mesma qualidade gerada pelo modelo, evitando qualquer degradação de qualidade.

4 Análise de Resultados

Por forma a avaliar as músicas geradas pelo modelo, o grupo optou por uma abordagem centrada na avaliação humana. Tendo em conta o reconhecimento da natureza subjetiva da apreciação musical, procuramos efetuar uma avaliação imparcial, envolvendo o maior número possível de pessoas neste processo. Para isto, foi criado um formulário, com recurso ao Google Forms, que contou com cerca de 40 participantes.

O formulário elaborado contou com 3 músicas distintas, entre elas 2 geradas pelo modelo e outra consistindo num excerto de uma música original. Cada participante foi requisitado a responder a uma série de critérios subjetivos para cada uma destas músicas, numa escala de 1 (Muito Mau) a 5 (Muito Bom). Os critérios envolvidos foram:

- **Qualidade do Áudio:** Clareza, mixagem e qualidade geral da produção musical;
- **Criatividade e Originalidade:** Singularidade, criatividade e grau de novidade da composição;
- **Estilo e Género:** Adequação ao estilo e género musical;
- **Harmonia:** Coesão e a fluidez das notas musicais;
- **Melodia:** Atratividade e memorabilidade das sequências melódicas;
- **Ritmo:** Dinâmica rítmica da música;
- **Emoção:** Capacidade da música de evocar sentimentos e emoções;
- **Origem:** Origem da composição da música, sendo as hipóteses a geração humana e a geração por inteligência articial.

Nas tabelas 1, 2 e 3 apresentam-se os resultados do formulário para cada música. De notar que as músicas geradas pelo modelo foram a música 1 e 2, enquanto a música 3 corresponde à música original, tendo sido retirada do Youtube.

	1	**2**	**3**	**4**	**5**	**Humano**	**IA**
Qualidade do Áudio	3	15	8	7	7	-	-
Criatividade e Originalidade	3	12	12	7	6	-	-
Estilo e Género	1	5	11	16	7	-	-
Harmonia	1	9	15	10	5	-	-
Melodia	4	5	14	11	6	-	-
Ritmo	0	10	15	7	8	-	-
Emoção	10	10	7	5	8	-	-
Origem	-	-	-	-	-	7	33

Tabela 1: Respostas relativas à musica 1.

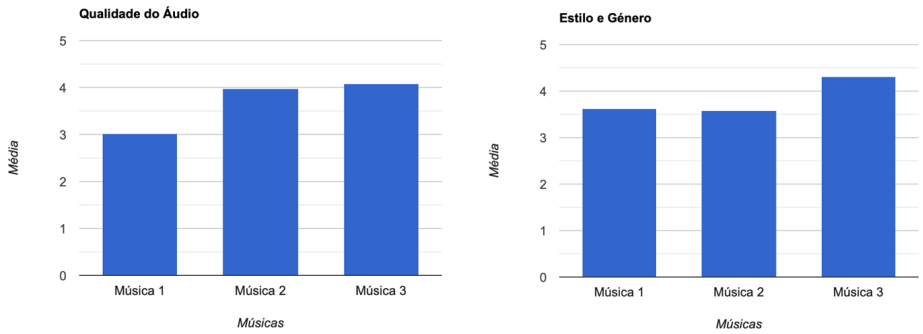
	1	2	3	4	5	Humano	IA
Qualidade do Áudio	0	2	9	17	12	-	-
Criatividade e Originalidade	0	1	17	15	7	-	-
Estilo e Género	1	5	9	20	5	-	-
Harmonia	0	1	7	20	12	-	-
Melodia	1	2	7	22	8	-	-
Ritmo	0	1	5	25	9	-	-
Emoção	1	5	11	17	6	-	-
Origem	-	-	-	-	-	15	25

Tabela 2: Respostas relativas à musica 2.

	1	2	3	4	5	Humano	IA
Qualidade do Áudio	0	2	8	14	16	-	-
Criatividade e Originalidade	0	4	13	11	12	-	-
Estilo e Género	0	0	5	17	18	-	-
Harmonia	0	1	5	16	18	-	-
Melodia	0	0	8	12	20	-	-
Ritmo	0	0	8	12	20	-	-
Emoção	1	3	6	17	13	-	-
Origem	-	-	-	-	-	25	15

Tabela 3: Respostas relativas à musica 3.

De maneira a sintetizar e compreender os dados anteriormente apresentados, realizamos diversos gráficos relativos às métricas “Qualidade do Áudio“, “Estilo e Género“ e “Origem“ para as 3 músicas apresentadas nos formulários.



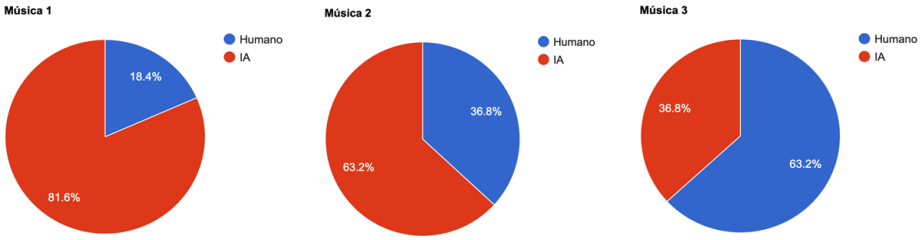
(a) Qualidade do Áudio.

(b) Estilo e Género.

Figura 1: Qualidade do Áudio e Estilo e Género das músicas.

Relativamente à qualidade do áudio das músicas, conseguimos concluir que a segunda música, gerada por inteligência artificial, alcançou uma média muito semelhante à música relativa a um áudio original de uma desgarrada. Apesar da música 1 ter obtido pior desempenho nesta métrica, não se distancia muito das médias das restantes músicas, o que demonstra a capacidade deste tipo de modelos de linguagem para gerar músicas.

No que diz respeito à métrica "Estilo e Género", a música 3 obteve a melhor classificação, algo expectável, visto que é um excerto de uma desgarrada verdadeira. Ainda assim, as músicas geradas pelo modelo obtiveram médias por volta de 3.5, algo bastante razoável para um modelo com bastantes limitações a nível computacional.



(a) Origem da música 1. (b) Origem da música 2. (c) Origem da música 3.

Figura 2: Origem das músicas.

A métrica da "Origem" visa medir a perceção dos questionados em distinguir se as músicas foram ou não geradas por inteligência artificial. De um modo geral, a maioria dos participantes conseguiu distinguir a origem da geração das músicas. No entanto, o grupo esperava que a música 3 tivesse uma maior percentagem de "Humano", algo que reflete uma das preocupações atuais que consiste na capacidade de distinguir algo que é gerado por humanos ou por inteligência artificial.

5 Limitações e Trabalho Futuro

Ao lidar com ambientes como o Google Colab, é importante estar ciente das limitações computacionais que podem surgir. Esta plataforma oferece acesso gratuito à GPU T4, para acelerar o processo de treino dos modelos. No entanto, existem várias restrições que impactam significativamente o desenvolvimento do projeto.

Primeiramente, como foi referido, o tempo de uso e os recursos computacionais disponíveis para cada sessão são limitados. Essa restrição implica a necessidade constante de criar novas contas para prolongar o acesso aos recursos, o que pode

ser um processo tedioso e ineficiente. Além disso, essas limitações impõem barreiras à complexidade dos modelos que podem ser treinados e ao tamanho dos conjuntos de dados que podem ser utilizados. Consequentemente, a capacidade de realizar otimizações ou treinar modelos em grande escala é severamente restringida, exigindo uma gestão cuidadosa dos recursos disponíveis para alcançar os melhores resultados possíveis.

Outra limitação significativa relaciona-se à quantidade e qualidade dos dados disponíveis para o treino. A falta de *datasets* específicos e amplos sobre o género de desgarrada foi um desafio. A principal fonte de dados utilizada, o YouTube, possui um número limitado de conteúdos relevantes, e muitos dos vídeos disponíveis apresentam qualidade de áudio inferior ao desejável. Isto não só limita a quantidade de músicas que podem ser utilizadas para treino, mas também pode comprometer a diversidade, a representatividade dos dados e a qualidade dos resultados.

Relativamente ao trabalho futuro, o próximo passo envolveria o *deployment* do modelo através de uma interface *web* intuitiva e funcional para garantir que o modelo desenvolvido fosse acessível e facilmente utilizável. Para realizar o *deployment* do modelo seria necessário um servidor para fazer a geração da música, o desenvolvimento de uma interface *web*, por forma a garantir a interação do utilizador com o modelo, e, finalmente, a integração do modelo, fazendo a ligação entre o *frontend* e o *backend* através de uma API REST, garantindo, assim, a possibilidade aos utilizadores de enviar *prompts* e receber as músicas geradas pelo modelo.

6 Conclusão

Apesar das diversas limitações mencionadas encontradas ao longo do projeto, como as restrições computacionais do Google Colab e a escassez de dados de qualidade disponíveis, a equipa foi capaz de treinar o modelo MusicGen para gerar músicas no género desgarrada de forma eficaz. Antes deste treino específico, o MusicGen não conseguia gerar músicas adequadas para este género, demonstrando a importância e o impacto do trabalho realizado.

Os questionários realizados para avaliar as músicas geradas obtiveram resultados satisfatórios. Embora a música original contida no questionário tenha obtido resultados superiores, o que já era esperado, as músicas geradas pelo modelo apresentaram resultados semelhantes e não foram substancialmente inferiores. Portanto, apesar das dificuldades, o modelo treinado conseguiu capturar características essenciais do género desgarrada, produzindo músicas que foram bem recebidas pelos avaliadores.

O projeto demonstrou ser possível treinar o MusicGen para um género musical específico, mesmo enfrentando limitações significativas de recursos e dados. Portanto, é possível supor que, com os recursos computacionais apropriados e com uma melhor e mais completa seleção dos dados de treino, seria possível gerar músicas com resultados muito mais satisfatórios.