Generación de imágenes a partir de letras de canciones

Basile Álvarez Andrés José

UNAM, FI Procesamiento del Lenguaje Natural 2023-2

16 de junio de 2023

Resumen

El presente proyecto de investigación propone utilizar redes neuronales adversariales en el procesamiento de lenguaje natural para generar imágenes que representen los sentimientos y emociones expresados en letras de canciones. La generación de imágenes a partir de texto requiere comprender y analizar el contenido emocional de las letras, y posteriormente sintetizar imágenes que evocan dichos sentimientos. Se utilizó un modelo DistilRoBERTa ajustado a clasificación de emociones y se generaron imágenes con redes neuronales adversariales entrenadas en subconjuntos del conjunto de datos WikiArt.

1. Introducción

La intersección entre el arte y la tecnología ha llevado al desarrollo de enfoques innovadores para la creación de contenido visual. En este trabajo de investigación, presentamos los resultados de un proyecto en el que se empleó una red neuronal adversarial (GAN, por sus siglas en inglés) para generar imágenes a partir de letras de canciones.

La generación de imágenes a partir de textos utilizando GAN es una tarea computacionalmente pesada que, como se menciona en (Chenshuang Zhang, 2023), últimamente ha sido reemplazada por modelos de difusión con similar peso computacional, pero mejores resultados. No obstante, el entrenamiento de una red neuronal de características similares a las presentadas en la *figura 1* para modelos de texto a imagen quedan fuera del alcance de nuestro poder computacional y de nuestros objetivos en este proyecto; por lo que decidimos seguir un esquema metodológico distinto (presentado en la metodología), que nos permita conocer más a fondo el funcionamiento básico de las GAN y de otros modelos de clasificación de textos y de generación de embeddings, en síntesis de los temas vistos en clase de Procesamiento de Lenguaje Natural.

El objetivo principal de nuestro proyecto fue explorar la capacidad de las GAN para generar imágenes que capturaran la esencia emocional de las letras de las

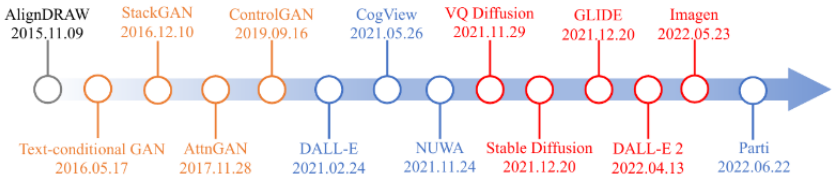

Figura 1: Modelos más importantes de generación de imágenes a partir de texto. Obtenido de (Chenshuang Zhang, 2023)

canciones. En el campo de la inteligencia artificial, se han desarrollado diferentes métodos y enfoques para abordar la tarea de análisis de emociones en textos, como el uso de Transformers (Ashish Vaswani, 2017) en BERT (Jacob Devlin, 2019) o GPT. Sin embargo, la aplicación de estos métodos al análisis de emociones en letras de canciones plantea desafíos adicionales debido a las particularidades artísticas y líricas de la música.

Para lograr un análisis adecuado, utilizaremos el modelo de lenguaje ajustado a clasificación de emociones, DistilRoBERTa-base. Categorizaremos las letras de las canciones en diferentes emociones, como tristeza, miedo o sorpresa. Este modelo pre-entrenado nos permmitirá obtener una representación numérica de los sentimientos asociados con cada letra de canción. Basado en esta clasificación, se creará un subconjunto de imágenes del conjunto de datos WikiArt para utilizarlo como parte del entrenamiento de la GAN. El subconjunto utilizado contendrá únicamente imágenes que previamente fueron clasificadas con la misma emoción que la letra de la canción y con una descripción con significado similar. En el apéndice se presenta más información sobre los conjuntos de datos y modelos utilizados durante esta investigación.

Si bien el presente trabajo de investigación no pretende ser una gran innovación en sí mismo, su objetivo radica en sentar las bases fundamentales para la construcción de modelos de inteligencia artificial más avanzados en el futuro. Reconocemos la importancia de explorar y comprender el análisis de emociones en letras de canciones y la generación de imágenes relacionadas, y consideramos que este proyecto es un paso inicial crucial en el camino hacia la mejora de los sistemas de inteligencia artificial en este ámbito. Al establecer los fundamentos teóricos y prácticos necesarios, esperamos allanar el camino para futuros avances y contribuir al desarrollo de modelos más sofisticados y efectivos en el campo del análisis emocional en el contexto musical.

2. Trabajo relacionado

A lo largo de los años, se han realizado diversos trabajos relacionados con el análisis de emociones en letras de canciones (y en textos de todo tipo) y la generación de imágenes basadas en esas emociones. Estos estudios han contribuido al avance de la inteligencia artificial y la computación afectiva en el ámbito musical. A continuación, se presentarán algunos trabajos relevantes:

Para la generación de imágenes, las Deep Convolutional GAN (DCGAN) son muy utilizadas. (Alec Radford, 2016) las introdujo en el año 2016, donde utilizando varios conjuntos de datos de imágenes, muestra evidencia convincente de que el modelo de DCGAN aprende una jerarquía de representaciones, desde partes de objetos hasta escenas completas, tanto en el generador como en el discriminador. Usaremos este tipo de GAN para la generación de imágenes.

También, existen muchos estudios que analizan la emoción que transmite la música a partir del análisis de archivos de audio. (Yang, 2021) explica la dificultad de clasificar canciones en categorías emotivas debido a las características de no exclusividad de las emociones, donde se puede tener una canción que pertenezca a más de una categoría. Este artículo introduce el algoritmo Artificial Bee Colony (ABC) para mejorar la retropropagación en la red neuronal. El valor de salida del algoritmo ABC se utiliza como peso y umbral de la red neuronal, encargándose de ajustar los pesos de forma óptima. Sus resultados presentan un menor error cuadrático en la clasificación de canciones en emociones que modelos de máquinas de soporte vectorial o N-vecinos cercanos. De forma similar, (Karen Rosero, 2022) utiliza varios modelos de redes neuronales distintos para la clasificación de canciones en emociones analizando características de audio, encontrando que las redes neuronales convolucionales (CNN) tuvieron la mayor precisión en la clasificación (78 %).

Los trabajos de (Goodfellow, 2014) y (Inkawhich, 2022) fueron fundamentales para el desarrollo de la GAN de este proyecto, siendo que el primero de ellos introdujo este tipo de modelo en 2014 y el segundo trabajo en DCGAN ejemplifica el funcionamiento de esta arquitectura con un conjunto de imágenes de caras de celebridades, que puede ser adaptado para funcionar con otros conjuntos de datos o incluso ser modificado para condicionar la salida de la red adversarial.

Estos trabajos y otros similares han sentado las bases para el análisis de emociones en letras de canciones (y textos en general), así como para la generación de imágenes asociadas. Sin embargo, aún existen desafíos y limitaciones en cuanto a la precisión y la cobertura emocional en estos enfoques, lo que indica la necesidad de continuar investigando y mejorando los modelos de inteligencia artificial en este ámbito. Además, existen pocos análisis que realicen una clasificación de emociones en canciones utilizando un acercamiento multimodal, donde se tomen en cuenta tanto características del audio, como la letra de la canción.

3. Metodología

Como fue mencionado anteriormente, la metodología empleada durante esta investigación une diversos conceptos de procesamiento de lenguaje natural con varios modelos de redes neuronales diferentes. Se puede describir en seis secciones principales: Generación de conjunto de datos de canciones; preprocesamiento y TF-IDF; Zero-Shot BERT para clasificación de canciones por emoción; obtención de embeddings de letras de canciones y de descripciones de pinturas; selección de canción y creación de conjunto de imágenes para entrenamiento de la GAN; y generación de imágenes con la GAN.

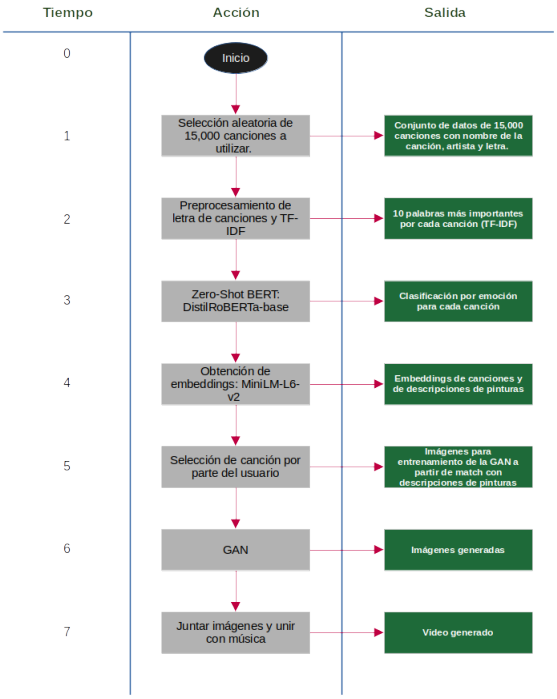

Figura 2: Secciones principales de la metodología

La **primera** de ellas utiliza el conjunto de datos compuesto de nombres de canciones, artistas y letra de las canciones para realizar una muestra aleatoria de 15,000 elementos, que serán los utilizados para el resto de la investigación. Esto se realiza con el fin de disminuir la cantidad de datos con la que se trabajará, ya que se cuenta con una capacidad computacional limitada.

En la **segunda sección**, se comienza por realizar el preprocesamiento del

conjunto de datos generado en la sección previa: Se eliminan los elementos vacíos, se convierte el texto a minúsculas, se aplican expresiones regulares para limpiar secuencias de caracteres como saltos de línea, números, entre otras acciones de preprocesamiento. Posteriormente, se eliminan las *stopwords* haciendo uso del conjunto de (NLTK, 2023). Finalmente, se utiliza la técnica TF-IDF para obtener las 10 palabras que más información aportan en cada canción y se almacena un conjunto de datos con dichas palabras.

Para la **tercera sección**, se utiliza el modelo DistilRoBERTa-base de (Hartmann, 2022), que fue previamente ajustado y entrenado para la clasificación de emociones y así obtener una predicción zero-shot de la emoción que mejor describe a cada una de las canciones. Las emociones en las que clasifica el modelo son: "Sadness", "Fear", "Disgust", "Anger", "Surprise", "Joy", y "Neutral", prácticamente las mismas que se tienen en las descripciones de pinturas, donde únicamente se cambió "Excitement" por "Surprise" y "Contentment" por "Joy". Fueron eliminadas las pinturas y canciones que expresabann "Disgust" debido a que se contaba con un número muy bajo de ellas.

En la **cuarta sección**, y con la idea de posteriormente tener una forma de asociar las palabras más importantes obtenidas con el TF-IDF de las letras de las canciones y las descripciones de las pinturas, se generaron embeddings para vincular los textos del TF-IDF (las diez palabras más significativas puestas en forma de oración) y las descripciones de pinturas con vectores de dimensión 384 utilizando el modelo (Huggingface, 2022). Este modelo fue especialmente creado para generar embeddings útiles para clusterización o búsqueda semántica, por lo que es útil para encontrar similitudes entre las palabras más significativas de las canciones y las descripciones de las pinturas.

La **quinta sección** permite seleccionar una canción dentro de la lista de 15,000 para calcular la distancia entre ésta y todas las descripciones de las pinturas que están catalogadas con la misma emoción que la canción. Por ejemplo, si una canción fue clasificada con DistilRoBERTa como "triste", entonces aquí se calculará únicamente la distancia entre la canción seleccionada y las descripciones de pinturas "tristes". Para limitar la salida de esta sección, se almacenarán únicamente las primeras 1500 descripciones de pinturas más similares al embedding de 10 palabras significativas de la letra de la canción. La salida de esta sección es un conjunto de datos que contiene las imágenes que serán ingresadas en la GAN para su entrenamiento.

La **sexta sección** contempla la creación de imágenes utilizando una GAN. A manera de introducción, se describe brevemente el funcionamiento de la DCGAN que se utilizó en esta investigación (Inkawhich, 2022). Una DCGAN es una extensión directa de la GAN descrita por Goodfellow, con la particularidad de que utiliza capas convolucionales y de convolución transpuesta de manera explícita en el discriminador y generador, respectivamente. En una GAN, *D(x)* representa la red discriminatoria, la cual genera la probabilidad de que una muestra x provenga de los datos de entrenamiento en lugar de ser generada por la red generadora G. En el caso de imágenes, la entrada de D(x) es una imagen de tamaño 3x64x64 (64 pixeles RGB). De manera intuitiva, D(x) debería ser alta cuando la muestra x sea de los datos de entrenamiento y baja cuando

provenga del generador. En esencia, D(x) funciona como un clasificador binario convencional.

En cuanto al generador, se utiliza z para representar un vector de espacio latente que se obtiene de una distribución normal estándar. *G(z)* es la función generadora que transforma el vector latente z en una muestra en el espacio de los datos. El objetivo del generador es estimar la distribución de los datos de entrenamiento (*Pdata*) para ser capaz de generar muestras falsas provenientes de esa distribución estimada (*Pg*).

Entonces, *D(G(z))* indica la probabilidad de que la salida del generador G sea una imagen real. En el marco conceptual de las GAN, se plantea un *minmax* entre el discriminador D y el generador G. D busca maximizar la probabilidad de clasificar correctamente muestras reales y falsas (logD(x)), mientras que G intenta minimizar la probabilidad de que D prediga que las salidas generadas son falsas, descrita como $\log(1 - D(G(z))$. La función de pérdida de la GAN se define como:

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[1 - \log D(G(z))] \tag{1}$$

En la DCGAN utilizada en la investigación, el discriminador está compuesto por capas de convolución con stride, capas de normalización por lotes (batch norm) y activaciones LeakyReLU. La entrada es una imagen de 3x64x64 y la salida es una probabilidad escalar de que la entrada provenga de la distribución de datos reales. El generador está compuesto por capas de convolución transpuesta, capas de normalización por lotes y activaciones ReLU. La entrada es un vector latente z extraído de una distribución normal estándar, y la salida es una imagen RGB de 3x64x64. Las capas convolucionales transpuestas permiten transformar el vector latente en un volumen con la misma forma que una imagen. La DCGAN utilizada puede verse en el diagrama de la figura 3.

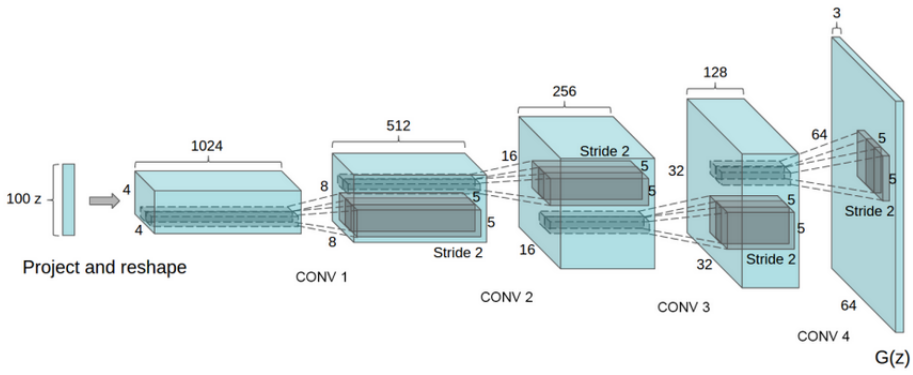

Figura 3: Diagrama de la DCGAN utilizada en esta investigación. Imagen de (Inkawhich, 2022).

Con la configuración de D y G, podemos especificar cómo aprenden ambas redes a través de las funciones de pérdida y optimizadores. Utilizaremos la función de pérdida de entropía cruzada binaria (BCELoss) que está definida como:

$$\ell(x, y) = L = \{l_1, \ldots, l_N\}^\top, \quad l_n = -\left[y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n)\right]$$

Establecemos que nuestra etiqueta para muestras reales es 1, mientras que la etiqueta para muestras falsas es 0. Estas etiquetas se utilizarán al calcular las pérdidas de los componentes discriminador (D) y generador (G), siguiendo la convención establecida en el artículo original de GAN. Además, configuramos dos optimizadores diferentes, uno para el discriminador y otro para el generador. Ambos optimizadores son del tipo Adam con una tasa de aprendizaje de 0.0002 y un valor de Beta1 de 0.5.

En la **séptima sección** de la metodología, se realizó un programa de Python que permite unir las imágenes generadas en la GAN y crear un video que pase entre ellas tratando de ir al ritmo de la canción escogida, definiendo el *frame rate* del video. La salida de esta sección es un video formato AVI, y se incluye un ejemplo de un video generado al finalizar la sección de análisis de este documento.

4. Experimentos

- Datos: Se emplearon varios conjuntos de datos durante la investigación. De éstos, se realizó un nuevo conjunto de datos que une información y se tomaron subconjuntos para el funcionamiento de nuestro sistema, como fue explicado anteriormente. Las principales fuentes de datos usadas son descritas a continuación:

  - MuSe Dataset: The Musical Sentiment Dataset (Akiki and Burghardt, 2021). Este conjunto de datos contiene información de 90,000 canciones. Incluye el nombre de la canción, el nombre del artista, y etiquetas de valencia (se refiere al grado de positividad o negatividad de una emoción), dominancia (se refiere al grado de control o influencia que una emoción ejerce sobre nosotros. Puede variar desde emociones que nos hacen sentir poderosos y en control, hasta emociones que nos hacen sentir sumisos o indefensos) y activación (se refiere al nivel de energía o intensidad que acompaña a una emoción). A partir de las etiquetas, se genera una columna que describe a la canción en una o más emociones. Para el trabajo en esta investigación, únicamente se tomó la primera etiqueta de emoción como la más importante y la que mejor describe a la canción.

- 150k Lyrics Labeled with Spotify Valence (Kaggle, 2022*a*): Este conjunto de datos fue creado para tareas de procesamiento de lenguaje natural, utilizando la etiqueta de valencia de Spotify como una medida de la positividad de cada canción. En esta investigación, se unió este conjunto de datos con MuSe para tener un único conjunto de canciones que incluya la letra, el nombre de la canción, el artista y la etiqueta de la emoción que transmite la canción.
- ArtEmis Dataset V2.0 (Mohamed et al., 2022): Este conjunto de datos contiene descripciones hechas por personas de miles de pinturas tomadas del conjunto de datos WikiArt. Además, contiene una columna que describe la emoción que transmite la pintura: "amusement", "anger", "awe", "contentment", "disgust", "emotion", "excitement", "fear", "sadness", "something else".
- WikiArt (Kaggle, 2022*b*): Conjunto de miles de imágenes de pinturas, catalogadas según su estilo y por artista. Se hizo la unión entre las descripciones de ArtEmis y las imágenes de WikiArt para seleccionar las imágenes de entrenamiento de la GAN.

- Método de evaluación: Al utilizar varios modelos distintos en nuestra investigación, se implementaron métricas de evaluación para conocer el funcionamiento de cada uno de ellos. La primera métrica de evaluación utilizada fue una matriz de confusión para conocer el porcentaje de etiquetas predecidas correctamente por DistilRoBERTa-base al momento de clasificar letras de canciones. En este caso en particular, la evaluación se realizó únicamente para las emociones "angry", "excitement" y "sadness", debido a que las etiquetas originales del conjunto de datos MuSe no eran exactamente las mismas que en las que clasifica el modelo DistilRoBERTa. No obstante, se supone que el modelo destilado de BERT fue específicamente entrenado para la clasificación de textos a emociones, como se describe en (Hartmann, 2022). Además de obtener la matriz de confusión, se obtuvieron métricas de precisión, recall y f1 score para conocer la calidad de la clasificación.

  La DCGAN fue evaluada utilizando las funciones de pérdida descritas en el apartado de metodología, y se calcularon conforme se entrenaba el modelo para entender de qué manera aprendían ambas redes neuronales. Se realizaron gráficas para medir la pérdida durante el entrenamiento y se evaluaron las imágenes generadas de forma cuantitativa, como se explica en la sección de análisis de este documento.

- Resultados: La clasificación de emociones utilizando DistilRoBERTa-base arrojó los siguientes resultados para las emociones “angry”, “excitement” y “sadness”:

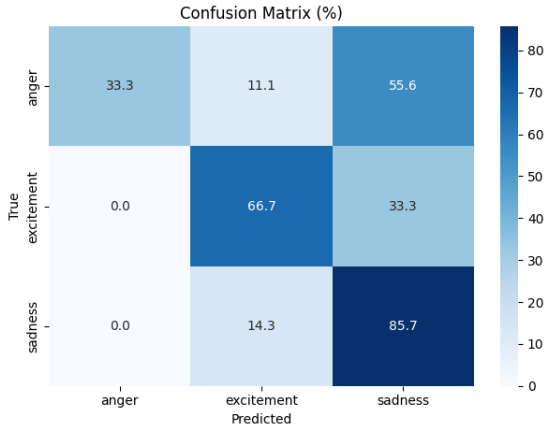

Figura 4: Matriz de confusión para la clasificación de canciones utilizando DistilRoBERTa-base Zero-Shot.

La precisión de la clasificación anterior fue de 76.92 %, con recall de 43.31 % y f1 score de 45.37 %.

Los resultados cuantitativos de la DCGAN pueden ser apreciados en las figuras 5 y 6, donde se grafica la pérdida en el generador y en el discriminador para dos canciones distintas. En los resultados se observa una gran pérdida al inicio del entrenamiento y una reducción paulatina conforme el generador y discriminador iban aprendiendo.

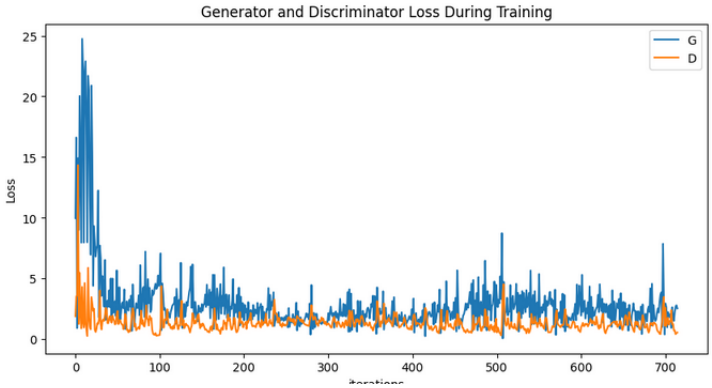

Figura 5: Pérdida en el Generador y Discriminador para The Long and Winding Road (Pérdida vs Iteraciones en 65 épocas).

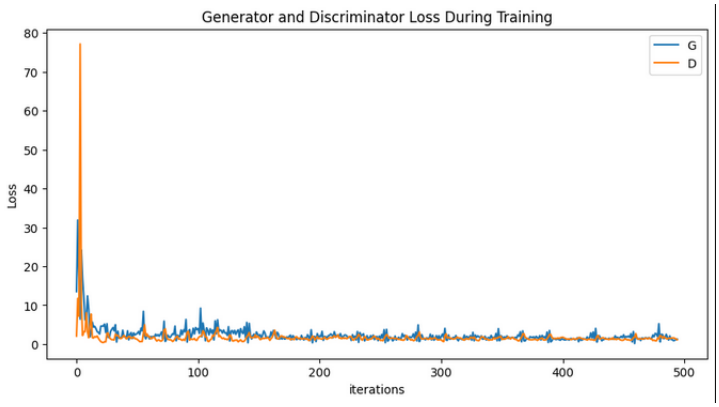

Figura 6: Pérdida en el Generador y Discriminador para Continental Drift.

5. Análisis

Como fue mencionado anteriormente, la clasificación utilizando DistilRoBERTa fue realizada Zero-Shot, lo cual quiere decir que no se realizó un ajuste a los pesos de la red neuronal, sino que se utilizó tal cual como estaba, sin entrenamiento posterior. No obstante, dicha red neuronal fue entrenada precisamente para clasificar en emociones, por lo que los resultados cuantitativos tuvieron una precisión aceptable. Observamos que el mayor error se encuentra al momento de distinguir entre enojo y tristeza. Esto puede deberse a que son emociones que se encuentran en el mismo nivel de valencia, sólo que con distintos niveles de activación, lo cual puede significar que la forma en la que las describimos en texto es similar, como se ve en la figura 7.

En la creación de embeddings y la obtención de la distancia entre las palabras

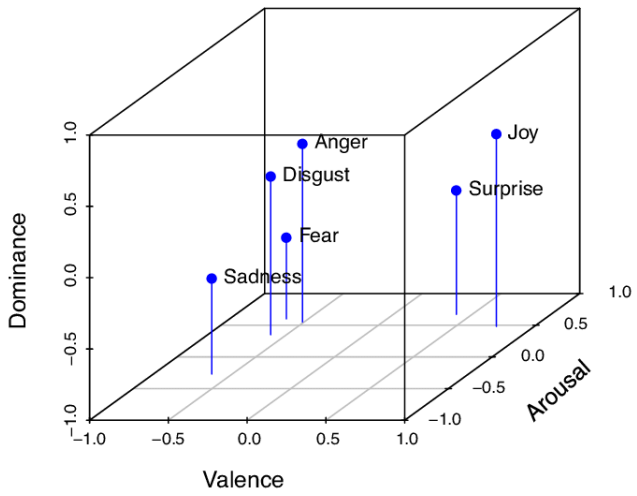

Figura 7: Valores de dominancia, valencia y activación para las emociones usadas en la clasificación en esta investigación. (Buechel and Hahn, 2018)

más significativas de la letra de una canción y las descripciones de las pinturas que evoquen la misma emoción que la canción, se observó que los resultados fueron mejores cuando las palabras más significativas de la canción eran poco comunes en las otras canciones y en las descripciones de las pinturas (por eso se trabajó con TF-IDF). Se presenta una ejemplificación de cómo palabras como "winding" o "lead" permiten encontrar las imágenes de pinturas más cercanas a la letra de "The Long and Winding Road".

En cuanto a los resultados de la GAN, se observó que muchas de las imágenes generadas presentan "colapso de moda". Este error se presenta cuando el generador de la GAN aprende a producir un resultado que engaña al discriminador, pero que repite ese resultado en muchas imágenes o en muchas secciones de una misma imagen. Cada iteración del generador se sobreoptimiza para un discriminador en particular, y el discriminador nunca logra aprender cómo salir de la trampa. Como remediación de este problema, los siguientes enfoques intentan forzar al generador a ampliar su alcance al evitar que se optimice para un único discriminador fijo:

- Pérdida de Wasserstein: La pérdida de Wasserstein alivia el colapso de moda al permitir entrenar al discriminador hasta la optimalidad sin preocuparse por los gradientes que desaparecen. Si el discriminador no queda atrapado en mínimos locales, aprende a rechazar las salidas en las que el generador se estabiliza. Entonces el generador tiene que intentar generar un tipo de imagen diferente.
- "Unrolled" GANs: Las "unrolled" GANs utilizan una función de pérdida del generador que incorpora no solo las clasificaciones del discriminador

actual, sino también las salidas de versiones futuras del discriminador. Así, el generador no puede sobreoptimizarse para un único discriminador.

Los enfoques anteriores podrían mejorar las salidas de la GAN. No obstante, también se tiene que considerar que, por capacidades computacionales, se trabajó con un conjunto de pequeño de imágenes de entrada a la GAN, se utilizaron pocas épocas (para no demorar demasiado tiempo entrenando) y se generaron imágenes de 64x64px, en las cuales es complicado distinguir detalles que tal vez generando imágenes más grandes con este mismo modelo sí se podrían alcanzar a percibir.

La DCGAN se utilizó para generar imágenes de dos canciones: "The Long and Winding Road" de "The Beatles" y "Continental Drift" de "The Rolling Stones". La primera de ellas fue catalogada como una canción "triste" por su letra, mientras que la segunda como una canción "excitante". Al analizar las imágenes generadas para cada una de ellas, vemos que las imágenes de la primera son en general más oscuras, con mucha utilización de azules y verdes oscuros, vinculados generalmente a la tristeza. En la generación de imágenes para la segunda canción, vemos que en general se trata de imágenes más claras, con abundancia de colores rojizos o naranjas, generalmente asociados con sensaciones que nos producen excitación.

También, al hacer un análisis cualitativo de las imágenes que se utilizaron como conjunto de datos de entrada para cada una de las canciones, vemos que las imágenes de "The Long and Winding Road" son en general más sombrías, y muestran escenas sin demasiado color o movimiento. En el caso del conjunto de imágenes de "Continental Drift", vemos imágenes con mayor movimiento y mayor uso de colores "vivos", lo cual se refleja en los colores utilizados en las imágenes generadas.

['winding', 'road', 'long', 'leads', 'many', 'ive', 'lead', 'stan', 'times', 'door']

[100]:

	art_style	painting	emotion	utterance	repetition
257594	Art_Nouveau_Modern	nicholas-roerich_the-oldest-the-wisest-1944	sadness	The man seems to be looking as if he has lost ...	10
185024	Abstract_Expressionism	jackson-pollock_mural	sadness	A tangled life often leads to a web of confusion.	5
202347	Minimalism	gene-davis_peach-glow-1958	sadness	the emptiness of it all, just lines and shadows	6
169225	Contemporary_Realism	lucian-freud_cyclamen-1964	sadness	Fading dying flowers.	5
149363	Post_Impressionism	henri-le-fauconnier_vielle-femme	sadness	Grandma hopes the Uber arrives before she turn...	5
...	...	...	...	...	...
224024	Pointillism	giacomo-balla_bankruptcy-1902	sadness	This door feels lonely and left behind	5
251336	Post_Impressionism	maxime-maufra_morning-in-winter-1905	sadness	This long and winding road looks very cold and...	5
294128	Realism	ivan-shishkin_the-road	sadness	Seeing this winding road lead to an abandoned-...	6
96629	Abstract_Expressionism	brice-marden_vine-1993	sadness	The lines are long but don't lead anywhere exc...	6
378522	Impressionism	william-merritt-chase_shinnecock-landscape-2	sadness	The road is so long that it looks like the jou...	5

1500 rows × 5 columns

Figura 8: Algunas de las pinturas seleccionadas para “The Long and Winding Road” de “The Beatles”. Estas fueron seleccionadas utilizando la técnica descrita en la quinta sección de la metodología. Se imprimen las 10 palabras más significativas de la letra de la canción (que fue catalogada como “triste” por DistilRoBERTa-base), y la multiplicación de su embedding con el embedding de las descripciones de pinturas tristes nos permite encontrar las pinturas que seleccionamos para el entrenamiento de la GAN. Es fácilmente observable que las descripciones de las pinturas son similares en significado a la letra de la canción.

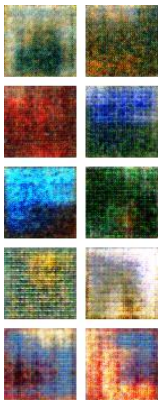

Figura 9: Algunas imágenes generadas para “The Long and Winding Road”.

Figura 10: Algunas de las imágenes utilizadas para el entrenamiento de la GAN en “The Long and Winding Road”.

Figura 11: Algunas de las imágenes utilizadas para el entrenamiento de la GAN en “Continental Drift”.

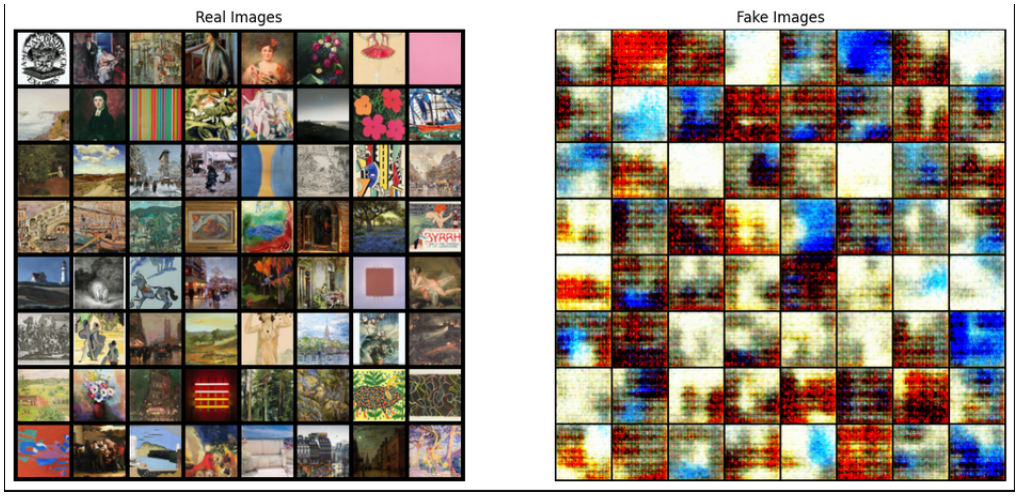

Figura 12: Algunas imágenes generadas para “Continental Drift”.

El video generado a partir de las imágenes de la DCGAN para la canción “The Long and Winding Road” de “The Beatles”, puede ser visto en el siguiente link. El video fue generado automáticamente, como fue descrito en la séptima sección de la metodología: https://drive.google.com/file/d/1t3M4Y_Y0kVe1hIPeRrYDDykbxVmeQb1y/view?usp=drive_link

6. Conclusiones

En conclusión, este proyecto de investigación ha abordado el desafío de clasificar letras de canciones en diferentes emociones y establecer una conexión entre las canciones seleccionadas por el usuario y las descripciones de pinturas que evocan sentimientos similares. A través del uso de técnicas de procesamiento de lenguaje natural y de la implementación de modelos avanzados (DistilRoBERTa-base y MiniLM-L6-v2) pre-entrenados, se logró categorizar las letras de canciones en distintas emociones de manera efectiva. Sin embargo, la precisión de la clasificación se podría mejorar al hacer ajustes a los pesos de la red neuronal al entrenar con nuestros datos.

Posteriormente, se exploró la vinculación entre las canciones y las descripciones de pinturas, buscando establecer una asociación visual con las emociones transmitidas por las letras. Al seleccionar imágenes de pinturas que evocan los mismos sentimientos, y que sus descripciones contienen palabras similares a las más importantes en la letra de la canción, se pudo generar un enlace entre texto de canciones y el arte visual.

Finalmente, se aplicó una Generative Adversarial Network (GAN) para entrenar un modelo capaz de generar nuevas imágenes basadas en las pinturas seleccionadas. Este enfoque permitió ampliar el repertorio visual y creativo, generando nuevas imágenes que capturan los sentimientos expresados tanto en las letras de las canciones como en las descripciones de pinturas.

En conjunto, este proyecto ha demostrado cómo la clasificación de letras de canciones, la asociación emocional con pinturas y el uso de GANs pueden converger en la creación imágenes únicas que busquen fusionar la música y la pintura. Los resultados obtenidos buscan ser fundacionales en cuanto a brindar los conocimientos y experiencia necesarios para poder abordar problemas de mayor complejidad, así como modelos más avanzados, en un futuro.

Como próximas líneas de trabajo, se piensa realizar una implementación de GAN condicionada, donde la entrada a la GAN sea directamente el texto de la canción. Además, se buscará crear un modelo de interpolación de imágenes para generar un video más fluido, así como responsivo a cambios en el audio de la canción (por ejemplo, que sonidos graves fuertes hagan “vibrar” la imagen). Otra línea de investigación futura consiste en la realización de un sistema similar, pero utilizando Stable Difussion.

Referencias

Akiki, C. and Burghardt, M. (2021), 'Muse: The musical sentiment dataset', *Journal of Open Humanities Data* **7**.

Alec Radford, L. M. (2016), 'Unsupervised representation learning with deep convolutional generative adversarial networks', *arXiv:1511.06434v2 [cs.LG] 7 Jan 2016* .

Ashish Vaswani, Noam Shazeer, N. P. J. U. (2017), 'Attention is all you need', *arXiv:1706.03762v5 [cs.CL] 6 Dec 2017* .

Buechel, S. and Hahn, U. (2018), Emotion representation mapping for automatic lexicon construction (mostly) performs on human level.

Chenshuang Zhang, Chaoning Zhang, M. Z. I. S. K. (2023), 'Text-to-image diffusion models in generative ai: A survey', *arXiv:2303.07909v2 [cs.CV] 2 Apr 2023* .

Goodfellow, I. (2014), 'Generative adversarial networks', *arXiv:1406.2661* .

Hartmann, J. (2022), 'Emotion english distilroberta-base', https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Huggingface, S. . (2022), 'Sentencetransformers: Minilm-l6-v2', https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Inkawhich, N. (2022), 'Dcgan tutorial', *Pytorch: https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html* .

Jacob Devlin, Ming-Wei Chang, K. L. K. T. (2019), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv:1810.04805v2 [cs.CL] 24 May 2019* .

Kaggle, E. . (2022*a*), '150k lyrics labeled with spotify valence', https://www.kaggle.com/datasets/edenbd/150k-lyrics-labeled-with-spotify-valence.

Kaggle, W. I. . (2022*b*), 'Wikiart: Visual art encyclopedia', https://www.kaggle.com/datasets/ipythonx/wikiart-gangogh-creating-art-gan.

Karen Rosero, A. N. (2022), 'Song emotion recognition: A performance comparison between audio features and artificial neural networks', *arXiv:2209.12045v1 [cs.SD] 24 Sep 2022* .

Mohamed, Y., Khan, F. F., Haydarov, K. and Elhoseiny, M. (2022), It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', Vol. abs/2204.07660.

NLTK (2023), 'Nltk documentation', *https://www.nltk.org/* .

Yang, J. (2021), 'A novel music emotion recognition model using neural network technology', *https://doi.org/10.3389/fpsyg.2021.760060* .