Cálculo De Similitud Entre Imágenes Artísticas Y Generadas Por Stable Diffusion Utilizando Redes Convolucionales Siamesas

Navil Pineda Rugerio, Diego Castro Elvira
Ingeniería en Inteligencia Artificial

Dr. Ricardo Ramos Aguilar

Dr. Jesús García Ramírez

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

ÍNDICE

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

1. Introducción

- Crecimiento de los modelos generativos de imágenes en el arte
- Surge la necesidad de evaluar la originalidad de obras generadas por IA.
- Desarrollar un método para comparar imágenes originales de arte y creadas por un modelo generativo, centrándose en medir su grado de similitud.

2. Planteamiento del Problema

Definición del Problema

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

- La IA genera obras artísticas comparables a las humanas, dificultando evaluar su autenticidad y originalidad.
- Comparar en conjunto los elementos que componen una obra es complejo, debido a las variaciones perceptibles en cada pieza.

Justificación

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

- Falta de modelos especializados para la generación y evaluación de similitud en pinturas de diferentes géneros artísticos.
- Se propone un enfoque que combina la generación *image-to-image* y la evaluación de su similitud usando redes convolucionales siamesas.
- Se desarrollará un conjunto de datos que incluya obras originales y versiones generadas por IA.

3. Objetivos

Objetivo General

Construir un modelo para calcular la similitud entre imágenes de pinturas artísticas y aquellas generadas por *Stable Diffusion XL Refiner 1.0* mediante el uso de descriptores de imagen y una red convolucional siamesa.

Objetivos específicos

- Crear un conjunto de datos.
- Seleccionar y extraer los descriptores de las imágenes.
- Crear y entrenar una red convolucional siamesa para calcular la similitud entre las imágenes.

4. Hipótesis

Hipótesis

Las redes neuronales siamesas permiten medir cuantitativamente el nivel de similitud compositiva y estilística entre pinturas artísticas originales y sus contrapartes generadas por *Stable Diffusion XL Refiner* 1.0, mediante la identificación y comparación de sus patrones visuales o semánticos.

5. Metodología

Fases de la metodología

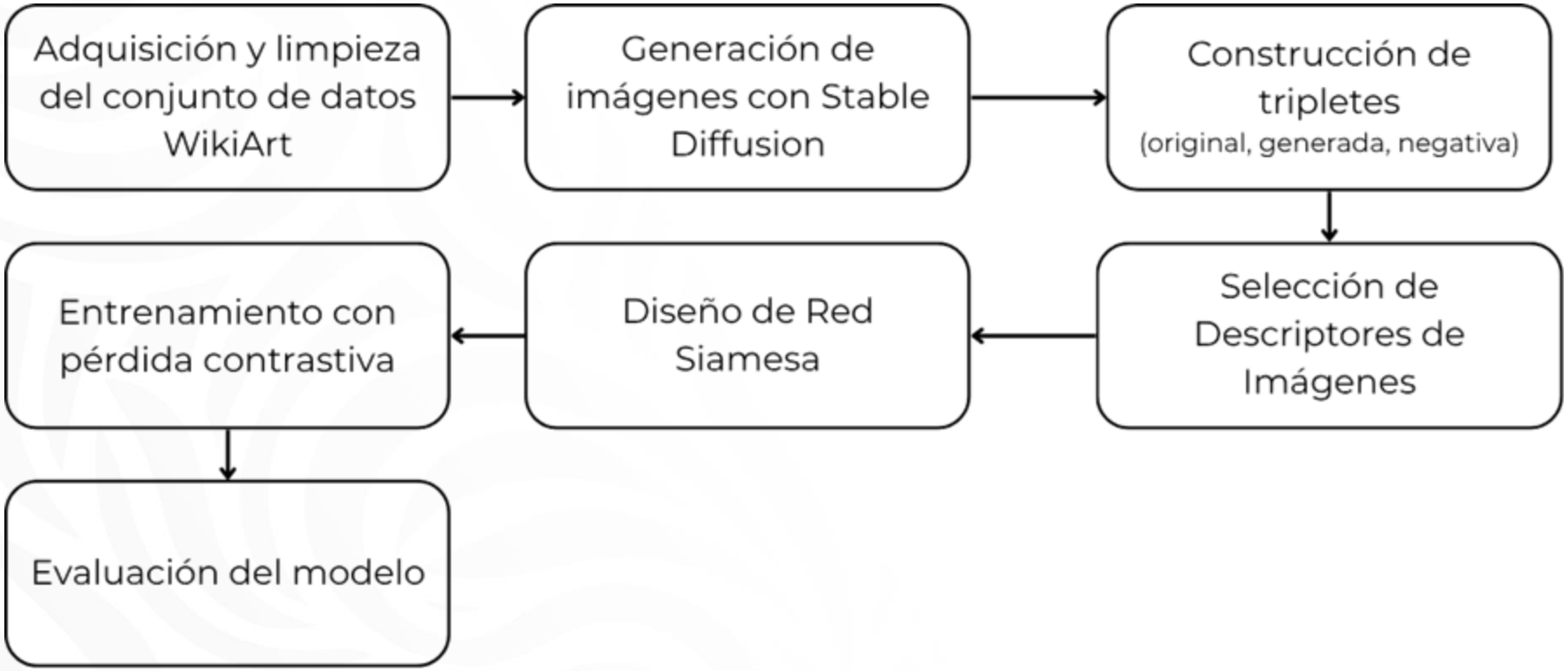


Figura 1. Metodología para la implementación de una Red Siamesa basada en CLIP. Fuente propia.

Adquisición del conjunto de datos

- En total se obtuvieron 81,445 imágenes, de los siguientes géneros artísticos.
- Cada imagen se asocia a un vector de metadatos compuesto por un subconjunto (entrenamiento o prueba), un género, y un artista.

#	Género Artístico	Cantidad
1	Impressionism	13060
2	Realism	10733
3	Romanticism	7019
4	Expressionism	6736
5	Post Impressionism	6450
6	Symbolism	4528
7	Art Nouveau Modern	4334
8	Baroque	4240
9	Abstract Expressionism	2782
10	Northern Renaissance	2552
11	Naive Art Primitivism	2405
12	Cubism	2235
13	Rococo	2089
14	Color Field Painting	1615
15	Pop Art	1483
16	Early Renaissance	1391
17	High Renaissance	1343
18	Minimalism	1337
19	Mannerism Late Renaissance	1279
20	Ukiyo e	1167
21	Fauvism	934
22	Pointillism	513
23	Contemporary Realism	481
24	New Realism	314
25	Synthetic Cubism	216
26	Analytical Cubism	110
27	Action Painting	98

Preprocesamiento de datos

- Interpolación a resoluciones de 1024×1024 y 768×768 píxeles.
- Redimensionado proporcional usando el algoritmo LANCZOS.
- Centrado con padding negro.

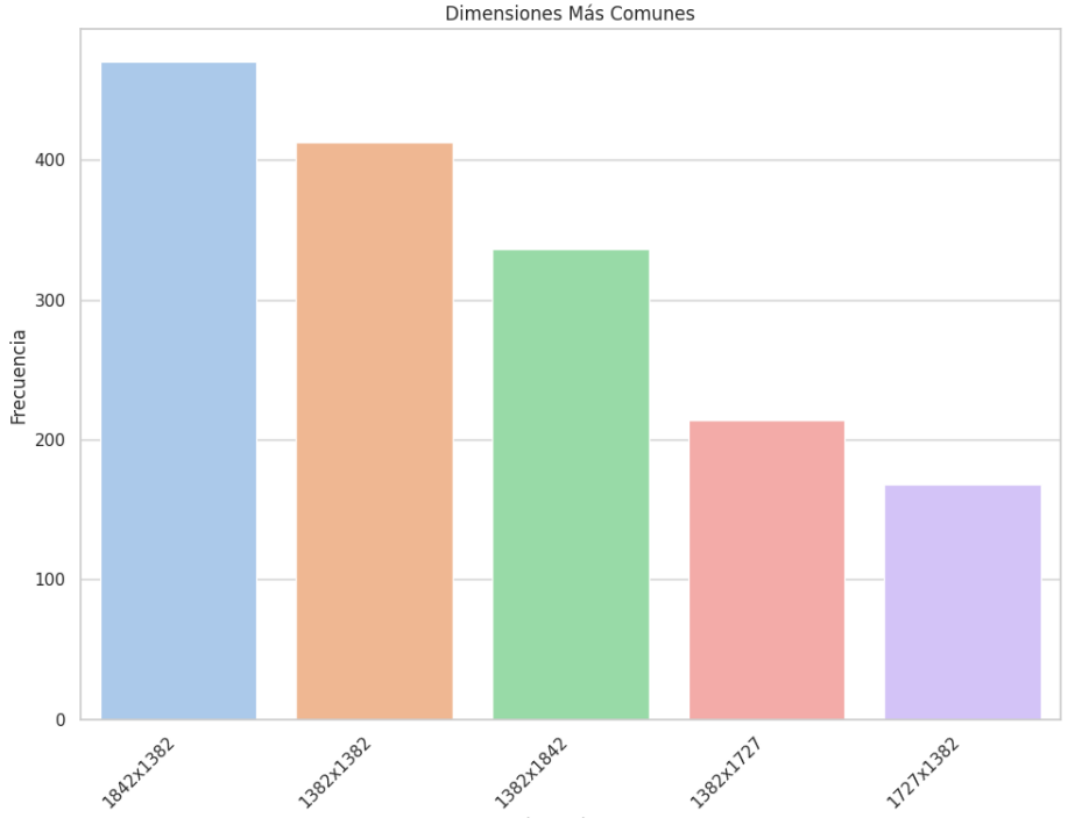


Figura 2. Dimensiones de imágenes más comunes en el conjunto de datos WikiArt.

Generación de imágenes con *Stable Diffusion*

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

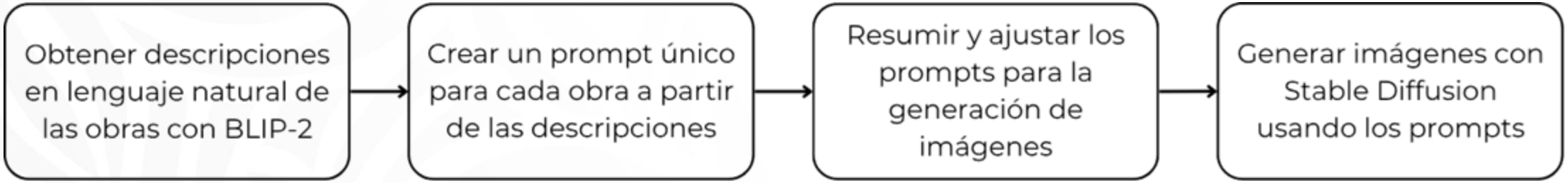


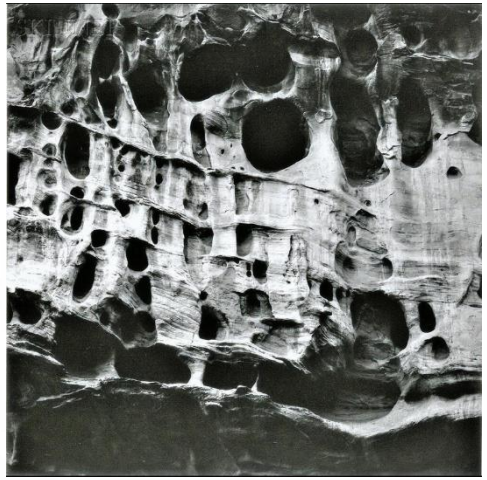
Figura 3. Fases para la generación de imágenes con Stable Diffusion. Fuente propia.

Obtención de descripciones con BLIP-2

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Obra: *antoine-blanchard_paris-les-champs-elysees*

Descripción con BLIP-2: Painting of people walking down the street in paris.



Obra: *aaron-siskind_utah-84-1976*

Descripción con BLIP-2: *A* black and white photo of a rock formation

Obra: *pyotr-konchalovsky_venice-palazzo-ducale-1924*

Descripción con BLIP-2: A painting of gondolas in front of a building

Figura 4. Ejemplos de descripciones generadas con el modelo BLIP-2.

Creación de *prompts* a partir de las descripciones

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

- Se diseñaron *prompts* para transformar las imágenes utilizando el modelo de Meta Llama-3-8B-Instruct-Lite.
- Se aplicaron dos niveles de transformación para diversificar el *dataset*:

Moderada	**Radical**
Variación de colores, cambios en luz o técnica artística, añadir o modificar detalles secundarios, mantiene el sujeto y composición original.	Transposición histórica, reinterpretación cultural, deformar la lógica visual, cambio de medio, reencuadre emocional, reinvención mitológica, estética tecnológica, cambio de escala, transformación simbólica.

Creación de *prompts* a partir de las descripciones

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Descripción con BLIP-2: *A painting of a white bird sitting on a vase*

Prompt radical

"A futuristic, neon-lit cityscape where the cockatoo is now a cybernetic being, perched atop a glowing, holographic vase" "The original's soft, dreamy atmosphere is replaced with a sense of high-tech urgency and innovation"

Descripción con BLIP-2: *A* painting of a woman holding a child

Prompt moderado

"A futuristic, neon-lit cityscape, where the woman's Victorian attire is replaced with a sleek, metallic jumpsuit" "The overall mood is one of high-tech futurism, with a hint of nostalgia for the past"

Figura 5. Ejemplos de prompts generados con el modelo Llama-3-8B-Instruct-Lite.

Ejemplos de Tripletas

1.

Imagen ancla I_a

Imagen positiva I_p

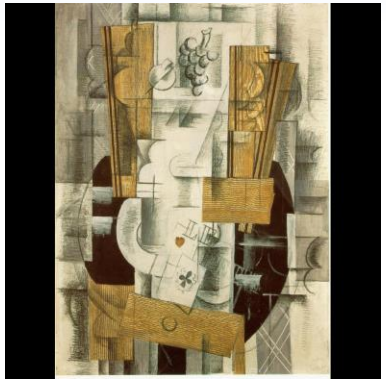


Imagen negativa I_n

2.

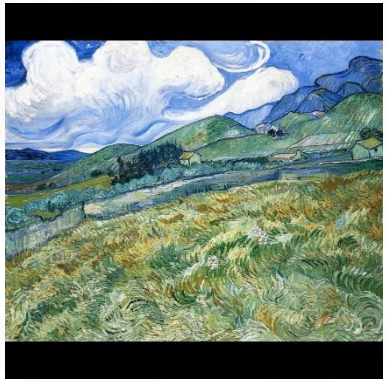


Imagen ancla I_a

Imagen positiva I_p

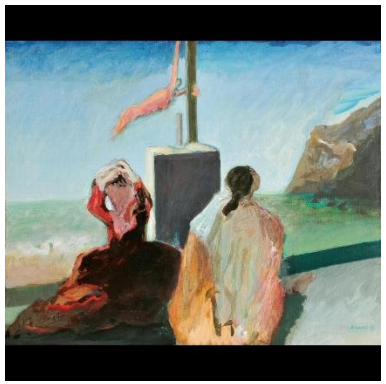


Imagen negativa I_n

Selección de modelo preentrenado

- Se seleccionó el modelo CLIP (clip-vit-base-patch32) por su capacidad para alinear representaciones visuales y textuales en un espacio común mediante aprendizaje contrastivo.
- El modelo de CLIP se mantiene congelado durante el entrenamiento aprovechando su capacidad de *zero-shot*.

Capas transformadoras

- 2 capas TransformerEncoder con:
  - Dimensión oculta: 1024
  - 8 cabezas de atención
  - Activación: GELU
  - Dropout: 0.1
- Proyección final con Linear + GELU + Dropout + Linear
- Salida normalizada para similitud del coseno

Entrenamiento

Para el entrenamiento, del conjunto de datos totales, se realizaron dos configuraciones para comparar el rendimiento del modelo:

1. **Configuración A:** 70% entrenamiento, 15% validación, 15% prueba.
2. **Configuración B:** 80% entrenamiento, 10% validación, 10% prueba.

Función de pérdida contrastiva basada en InfoNCE, que incentiva mayor similitud entre la imagen original y la positiva, y menor entre la original y la negativa.

$$Loss = -log(\frac{exp(\frac{sim_p}{\tau})}{(exp(\frac{sim_p}{\tau}) + exp(\frac{sim_n}{\tau})})$$

Figura 6. Función de pérdida contrastive basada en tripletes.

Evaluación

La métrica central utilizada para la comparación de *embeddings* fue la similitud coseno, definida como:

$$s(v_i, v_j) = \frac{v_i \cdot v_j}{\| v_i \| \| v_j \|} \in [1, 1]$$

Ecuacion X. Fórmula de Similitud Coseno.

Donde v_i y v_j son vectores de características extraídos por la Red Siamesa y normalizados previamente. A cada tripleta se le asignan dos puntuaciones de similitud.

Diseño de Red Siamesa

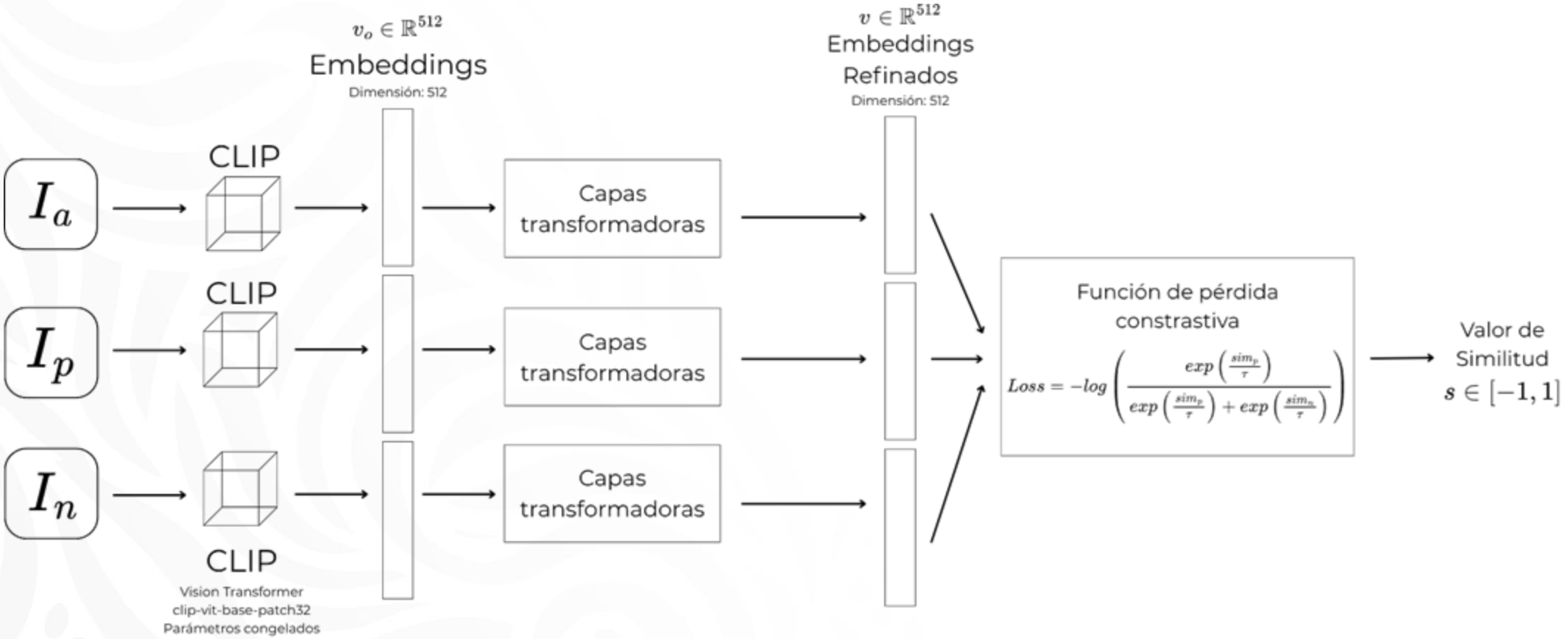


Figura 8. Modelo de Red Siamesa. Fuente propia.

6. Resultados

Función de pérdida de entrenamiento

La evolución de la función de pérdida para el conjunto de entrenamiento va de (≈0,17) y desciende de manera rápida durante las primeras cinco épocas, seguido por una reducción más suave, convergiendo hacia un valor cercano a 0,02.

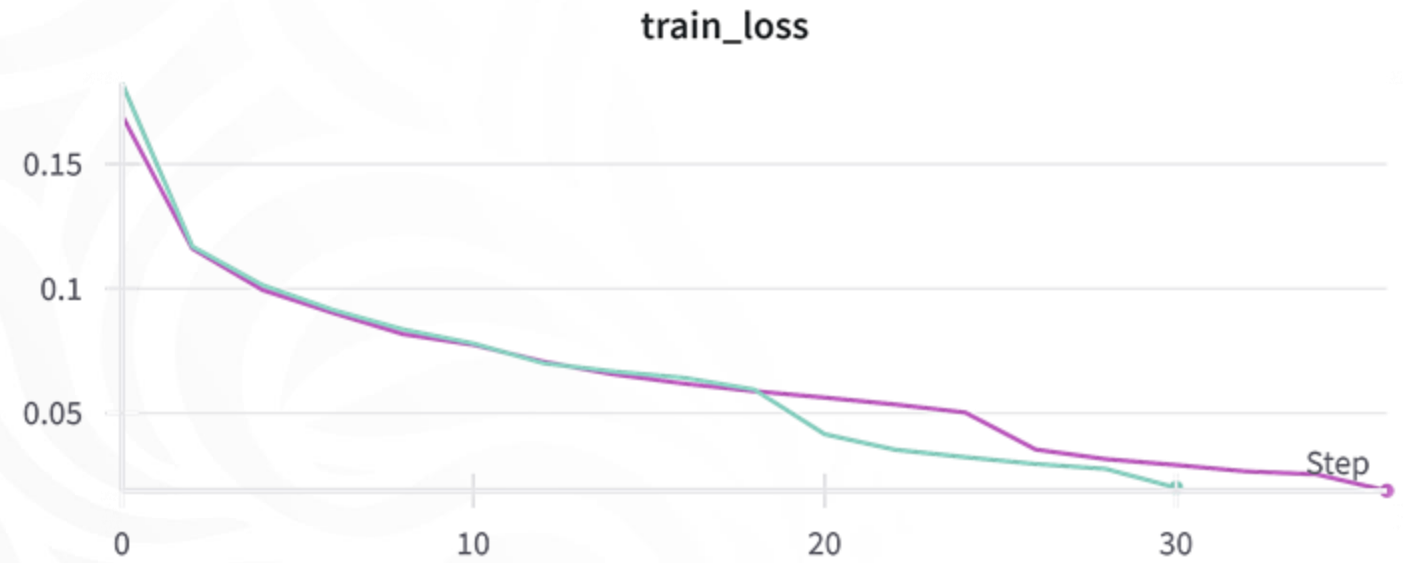


Figura 9. Función de pérdida de entrenamiento para las dos configuraciones del modelo. Fuente propia.

Función de pérdida de validación

Figura 10. Función de pérdida de validación para las dos configuraciones del modelo. Fuente propia.

Precisión de tripletas de entrenamiento

Respecto a la precisión sobre tripletas, la métrica *train_triplet_accuracy* parte de un valor inicial elevado (≈97,5%), y se observa una mejora acelerada durante las primeras épocas, alcanzando valores superiores al 99% y estabilizándose finalmente alrededor del 99.95%.

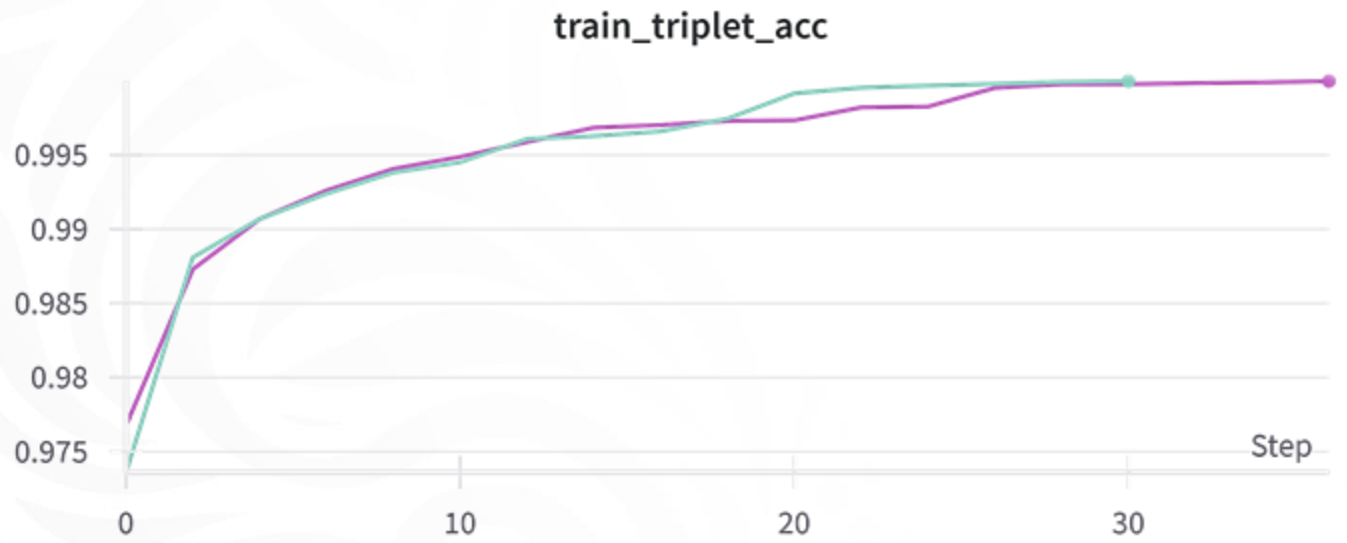


Figura 11. Precisión de tripletas de entrenamiento para las dos configuraciones del modelo. Fuente propia,

Precisión de tripletas sobre el conjunto de prueba

Se evaluaron las dos configuraciones del modelo, y los resultados cuantitativos se resumen en la siguiente tabla:

Configuración	Triplet Accuracy	μ_{pos}	μ_{neg}	$\Delta\mu = \mu_{pos} - \mu_{neg}$
70-15-15	**0.994**	**0.824**	**0.146**	**0.677**
80-10-10	0.985	0.848	0.385	0.463

Histograma de similitudes

En la Figura ocurre exactamente lo que se esperaba, las similitudes positivas se agrupen cerca de 1, mientras que las negativas tiendan hacia valores más bajos, cercanos a 0. La superposición mínima entre ambas curvas indica que si se logra una distinción entre ambas similitudes.

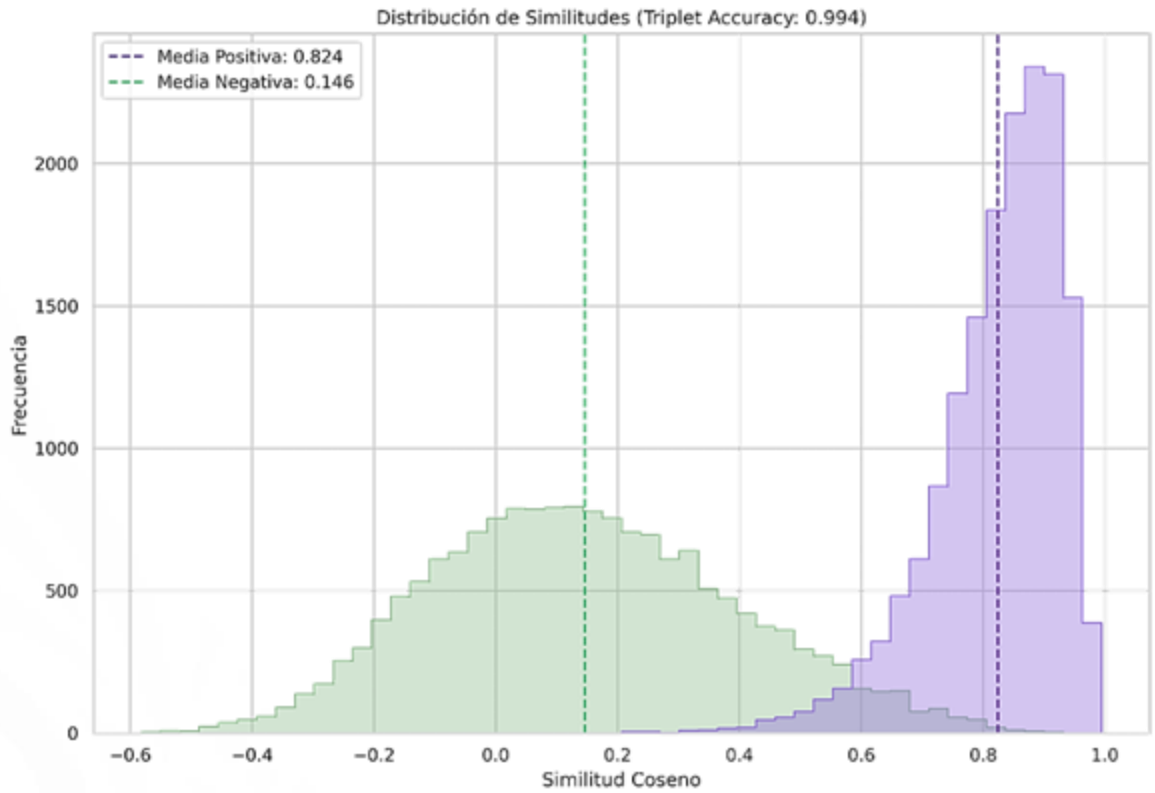


Figura 12. Histograma de Similitudes para el mejor modelo entrenado. Fuente propia.

Resultados de Inferencia

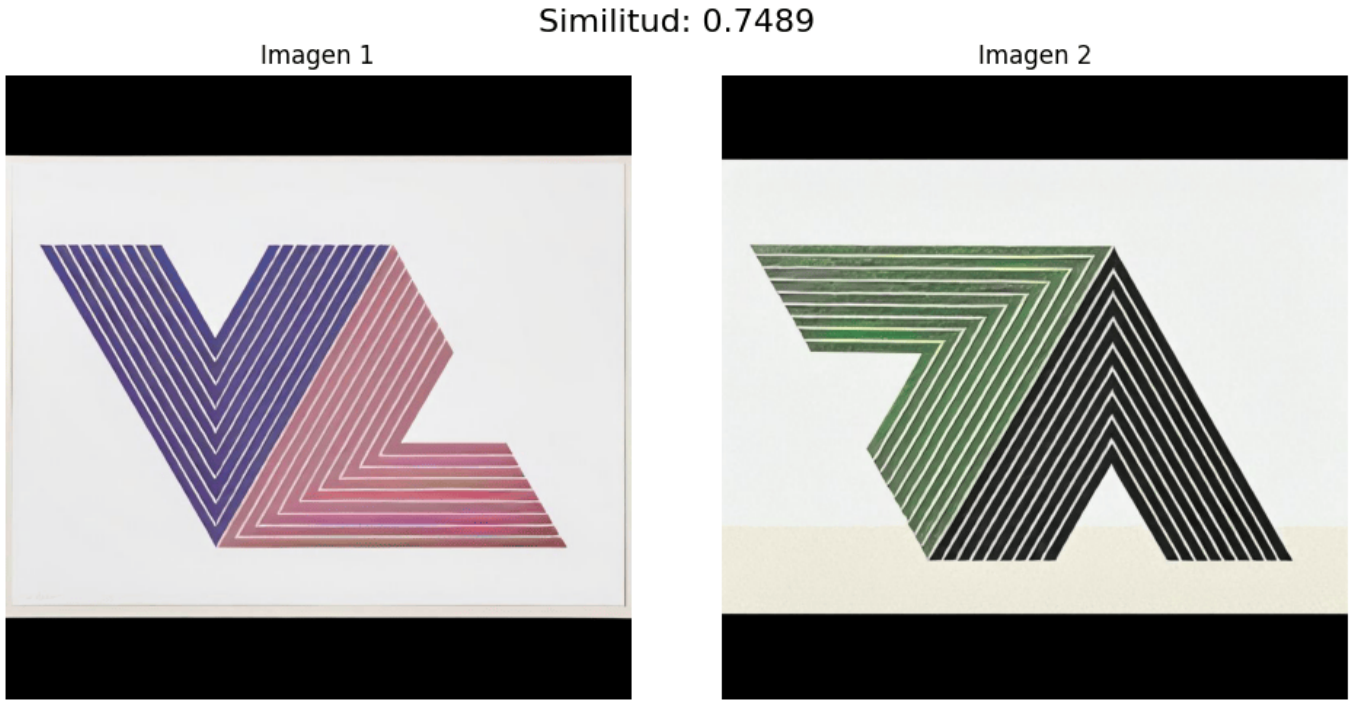


Figura 13. Resultado en tiempo de inferencia

Resultados de Inferencia

Figura 14. Resultado en tiempo de inferencia

Resultados de Inferencia

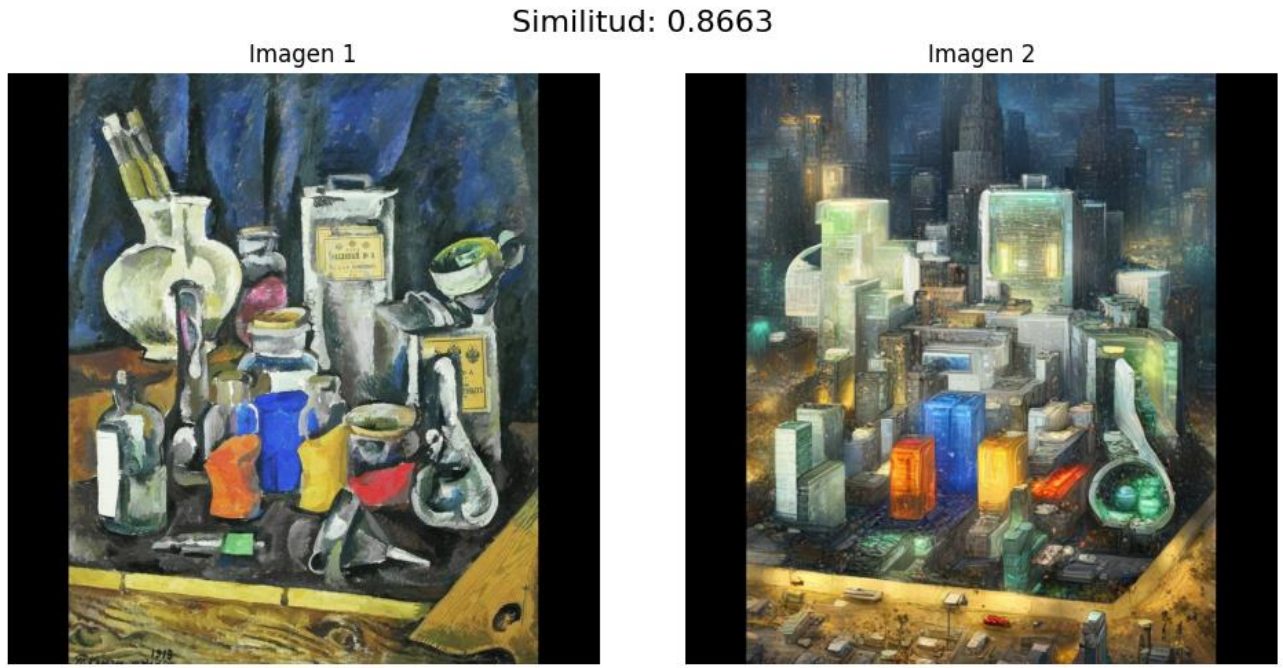


Figura 15. Resultado en tiempo de inferencia

Conclusiones

- Se entrenó una red siamesa con embeddings basados en CLIP y pérdida triplete para medir similitud semántica entre imágenes artísticas y generadas, logrando una precisión del 99.34% con clara separación entre pares positivos y negativos.
- Se demostró que el aprendizaje por similitud, sin clases explícitas, puede capturar patrones artísticos significativos, aunque se detectó un ligero sobreajuste hacia el final del entrenamiento.
- El trabajo resalta el potencial de modelos multimodales para verificar la autenticidad de imágenes generadas por IA, una herramienta relevante ante el aumento de réplicas estilísticas no autorizadas en industrias creativas.

Referencias

O'Shea, K. and Nash, R. (2015). An Introduction to Convolutional Neural Networks. Article arXiv:1511.08458v2. https://doi.org/10.48550/arXiv.1511.08458

Alshalali, T., & Josyula, D. (2018). Fine-tuning of pre-trained deep learn-ing models with extreme learning machine. 2018 International Con-ference on Computational Science and Computational Intelligence(CSCI), 469–473. https://doi.org/10.1109/CSCI46756.2018.00096

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning forimage recognition. 2016 IEEE Conference on Computer Vision andPattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

Koch, G. R. (2015). Siamese neural networks for one-shot image recog-nition. https://api.semanticscholar.org/CorpusID:13874643

Sherly A P, P. A. R. (2024). Siamese augmented network (saugnet) forjpeg steganalysis. https:// doi.org/ 10.36227/ techrxiv.171994747.74344541/v1

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller,J., Penna, J., & Rombach, R. (2023). Sdxl: Improving latent diffusionmodels for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.

Zhang, Y., Teoh, T. T., Lim, W. H., Wang, H., & Kawaguchi, K. (2024). OnCopyright Risks of Text-to-Image Diffusion Models. National Univer-sity of Singapore.

Du, L., Zhu, Z., Chen, M., Ji, S., Cheng, P., Chen, J., & Zhang, Z. (2024).WIP: Auditing Artist Style Pirate in Text-to-image Generation Models.CISPA Helmholtz Center for Information Security, Saarbrucken.

Casper, S., Guo, Z., Mogulothu, S., Marinov, Z., Deshpande, C., Yew, R. J.,Dai, Z., & Hadfield-Menell, D. (2023). Measuring the Success of Diffu-sion Models at Imitating Human Artists. International Conference onMachine Learning

Moayeri, M., Basu, S., Balasubramanian, S., Kattakinda, P., Chengini,A., Brauneis, R., & Feizi, S. (2024). Rethinking Artistic Copyright Infrin-gements in the Era of Text-to-Image Generative Models. University ofMaryland, Computer Science Department.

GRACIAS

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"