Generando Animes con *Stable Diffusion*

Carlos Cortés-Méndez
carlos.mendez@cimat.mx

Jean-Bernard Hayet
jbhayet@cimat.mx

7 de octubre de 2023

1. Motivación

En estos últimos años, los modelos generativos de inteligencia artificial han captado el interés del público general debido a sus impresionantes resultados. Permiten en particular generar datos de forma aleatoria, correspondiendo a una distribución de probabilidad meta que ha proveído muestras de entrenamiento. En particular, en el ámbito de visión computacional (para tareas involucrando generar imágenes), el nombre de *Stable Diffusion* es ahora ampliamente conocido [1, 2], ya que esta técnica ha servido de motor a los más espectaculares generadores. Sin embargo, muchos de estos resultados son producto de modelos comerciales, los cuales no publican sus hallazgos, avances o modelos.

En este trabajo en progreso, proponemos un esquema en donde un modelo de difusión con variable latente se utiliza para generar imágenes de caras de tipo anime/manga. Este esquema se puede usar de forma incondicional o condicional (pasando un esbozo de dicha cara) y podría ser útil en la generación automática de avatares, por ejemplo.

2. Metodología

Hemos seguido el procedimiento usual de los modelos de difusión con variable latente, en el cual se entrena primero un *autoencoder*, el cual permite obtener una representación compacta en un espacio de menor dimensión de las imágenes (el espacio latente). Esta red puede ser un autoencoder variacional (VAE) [3] o un autoencoder de vectores cuantizados (VQ-VAE) [4]. En nuestra aplicación, utilizamos vectores cuantizados, por haber sido mostrado más efectivo para reconstruir detalles.

Posteriormente, se entrena un modelo de difusión latente [5], el cual aplica un proceso de difusión *en el espacio latente* de menor dimensión generado por el autoencoder. Esto permite acelerar el proceso de difusión y generalizarlo para imágenes de mayor resolución. Recordamos que los modelos de difusión se entrenan con versiones ruidosas de imágenes originales, con un nivel de ruido progresivo y controlado. El entrenamiento consiste a aprender a estimar el ruido aplicado a la imagen original [2].

Para entrenar y evaluar nuestro método, hemos usado el conjunto de imágenes de anime `danbooru2021` [6], particularmente el subconjunto de caras.

3. Resultados

3.1. VQVAE

Nuestro mejor modelo actual es de 52 millones de parámetros, tomando aproximadamente 12 horas para entrenarse en 2 gpus TITAN RTX y alcanzando un PSNR de 25.31 y un LPIPS [7] de 0.1005. Algunos resultados ejemplo de la reconstrucción se observan en la figura 1, se observa que tiene algunos problemas con los detalles pero en general rendimiento es muy bueno.

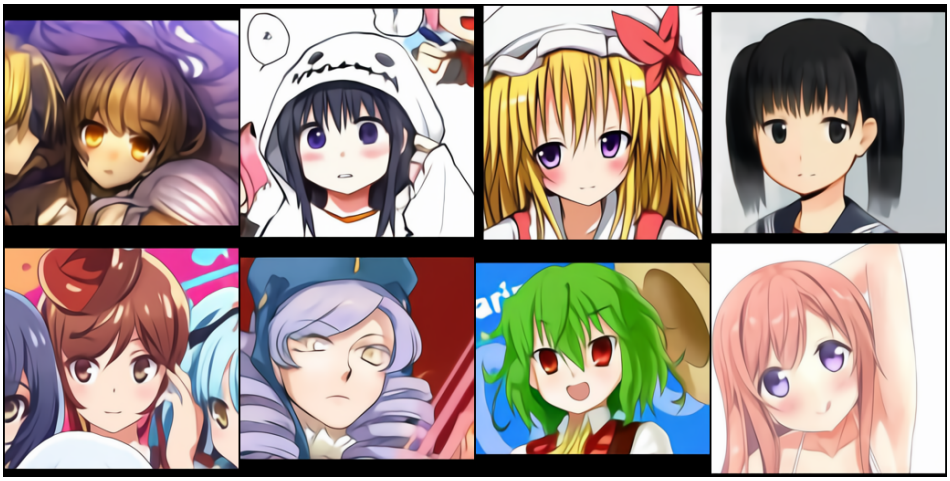

Figura 1: Imágenes reconstruidas usando el VQ-VAE

3.2. Difusión latente

Nuestro mejor modelo actual es de 80 millones de parámetros, tomando aproximadamente 24 horas para entrenarse en 2 gpus TITAN RTX. Algunos resultados ejemplo de las imágenes generadas se observan en la figura 2, se observa que algunas imágenes se ven muy realistas mientras que otras aun tienen mucho ruido.