

Image Edition Using Diffusion Models and Natural Language

María Graciela Cruz Cáceres

July 2023

Definición del problema

La edición de imágenes con modelos de difusión es una técnica que permite editar imágenes utilizando lenguaje natural.

Todavía presenta algunas limitaciones como:

- Limitaciones de **lenguaje natural**
- Necesidad de una gran cantidad de datos
- Limitaciones de **hardware**
- Dificultad para **conservar las características** originales

Objetivos

- Realizar el **levantamiento del estado del arte** en la edición de imágenes a partir del lenguaje natural utilizando modelos de difusión
- Identificar las **limitaciones** presentes en las técnicas más relevantes.
- Analizar y **comparar los resultados** obtenidos por ambas técnicas en diferentes métricas de evaluación como: la similitud entre las representaciones visual-textual y la similitud de la dirección de cambio imagen-texto entre la imagen original y la edición

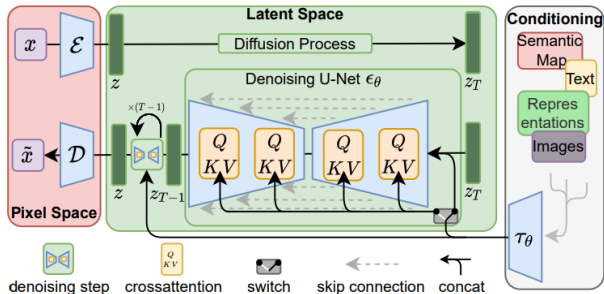
Estado del arte

Técnica	Información que conserva	Reentrenamiento y ajuste	Enfoque de Instrucción
Diffusionclip [1]	Semántica y estructural	Afinación del modelo	Instrucción a Instrucción
Prompt-to-Prompt [2]	Estructural	No requiere	Instrucción a Instrucción
Imagic [3]	Semántica	Afinación del modelo	Instrucción simple
Text2live [4]	Estructural	Reentrenamiento	Característica simple
Plug and Play [5]	Estructural	No requiere	Instrucción simple
Sine [6]	Semántica y estructural	Afinación del modelo	Característica simple
Pix2pix Zero [7]	Estructural	No requiere	Característica simple
HiPer [8]	Semántica	No requiere	Instrucción a Instrucción
MDP [9]	Estructural	No requiere	Instrucción a Instrucción

Table: Técnicas basadas en modelos de difusión para la edición de imágenes

Latent Diffusion Model LDM

- **Modelos de difusión** Son modelos probabilísticos que aprenden la distribución de los datos mediante la eliminación gradual de ruido de una variable distribuida normalmente, pueden ser interpretados como una secuencia de autoencoders de eliminación de ruido.
- Espacio latente eficiente y de baja dimensionalidad en el cual los detalles de alta frecuencia e imperceptibles se abstraen.



Plug-and-Play

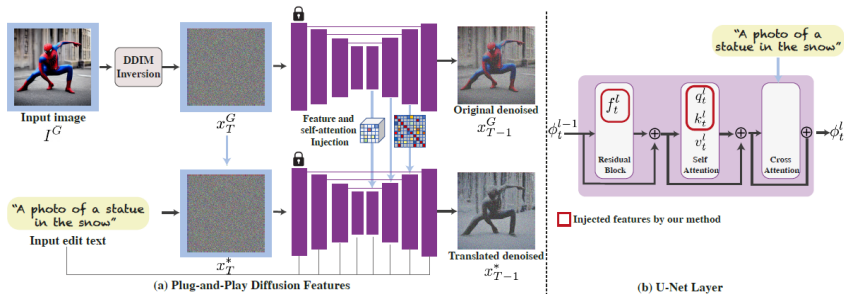


Figure: Características de difusión de Plug-and-Play. (a) El marco de trabajo toma como entrada una imagen de guía y una frase de texto que describe la traducción deseada. La imagen de guía se invierte para obtener un ruido inicial, que luego se va depurando progresivamente utilizando muestreo *DDIM*. Durante este proceso, extrae las características espaciales de las capas decodificadoras y su auto-atención, como se ilustra en (b). Para generar nuestra imagen traducida guiada por texto inyecta las características de guía en ciertas capas.

Hiper

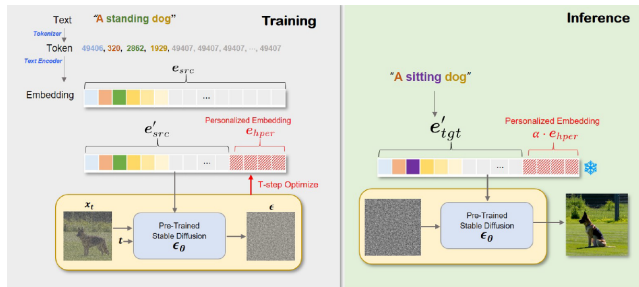


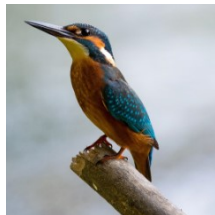
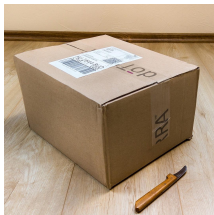
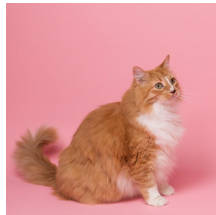
Figure: (Entrenamiento) Primero, se convierte la frase de texto de origen, en incrustación de texto. La parte informativa de la incrustación de destino y la incrustación personalizada se concatenan y se utilizan como entrada de la red neuronal pre-entrenada. Durante el entrenamiento, solo se optimiza la incrustación personalizada. (Inferencia) La incrustación de destino también se recorta y se concatena con la incrustación personalizada. El modelo pre-entrenado de texto a imagen, genera una imagen que tiene el significado del texto de destino y el sujeto de la imagen de origen.

Experimentos

- **Benchmark:** *TEdBench (Textual Editing Benchmark)*
- Técnicas implementadas:
 - **Plug and Play:** Wild TI2I
 - **Hiper:** TEdBench
- **Métricas:**
 - CLIP Score: Mide la similitud entre las representaciones visual y textual
 - Similaridad de Dirección Texto-Imagen: Mide la similitud entre la dirección del cambio entre dos leyendas imágenes
- **Ejecución:** Google Colab

Benchmark: TEdBench

- Propuesto por **Imagic** [3]
- Contiene 100 pares de imágenes-texto
- Texto: Edición simple de texto. Cambio de posturas y edición de múltiples objetos.
- Imagen: Imágenes reales fotos de animales y objetos, $1024 \times 1024px$ a color
- Ejemplos:
 - A photo of a jumping dog.
 - A photo of a sleeping cat.
 - A photo of an open box.
 - A photo of a bird spreading wings..



CLIP Score

Mide la similitud entre las representaciones visual y textual.

$$CLIPScore(I, C) = \max(\cos(E_I, E_C), 0) \quad (1)$$

Donde:

- E_I es la incrustación de una imagen I
- E_C es la incrustación de un texto C

El puntaje está limitado entre 0 y 1

Similaridad de Dirección Texto-Imagen de CLIP

Mide la consistencia del cambio entre las dos imágenes (en el espacio CLIP)

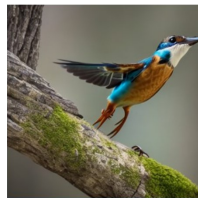
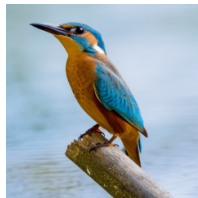
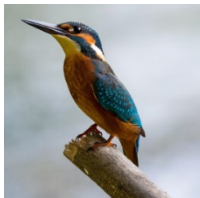
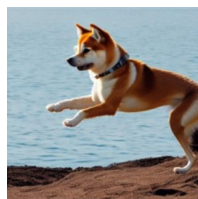
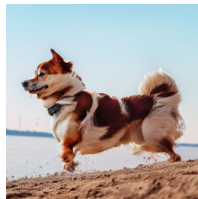
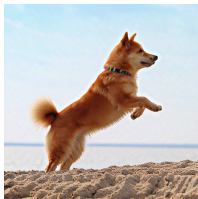
$$L_{direction} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I| \cdot |\Delta T|} \quad (2)$$

Donde:

- ΔT Diferencia en el espacio CLIP entre las dos imágenes
- ΔI Diferencia en el espacio CLIP entre las dos leyendas
- $L_{direction}$ es la métrica de consistencia que mide la relación entre las diferencias en las imágenes y las diferencias en las leyendas.

Experimentos y resultados

Se muestran la imagen original y la edición realizada por *Imagic*, *Plug and Play* e *Hiper*



Resultados de las métricas





Técnica	<i>CLIP Score</i>	<i>CLIP Similarity</i>
<i>Plug and Play</i> [5]	0.2245	0.2430
<i>Hiper</i> [8]	0.2173	0.1547
<i>Imagic</i> [3]	0.2148	0.1021

Table: Resultados de los puntajes obtenidos al editar imágenes con las técnicas *Plug and Play* e *Hiper* utilizando las métricas de *Clip Score* y *Clip text similarity*. Adicionalmente se presenta junto con las métricas obtenidas con la técnica *Imagic*






Conclusiones

- Existe un gran interés en el tema de edición de imágenes que se puede observar gracias a la **publicación continua** de estudios relacionados al tema
- Los **resultados de las métricas seleccionadas es bajo**. Es posible que no se hayan encontrado las métricas apropiadas y que existe aun espacio de mejora
- La técnica de *Plug and Play* obtuvo mejores resultados a comparación de *Hiper* y necesita menor cantidad de recursos para ser ejecutada.
- La diferencia entre las métricas no es muy alta por lo que no se puede afirmar que una es mejor que la otra
- La diferencia principal entre ambas técnicas es con respecto a su enfoque.
- Una de las principales limitaciones se encuentra la **limitación de recursos**.
- Algunos **espacios de mejora** son la preservación de la estructura en general y la composición espacial.
- Como posible trabajo futuro, se plantea **abordar la dificultad de especificar con lenguaje natural ediciones espaciales**.

Referencias I

-  G. Kim, T. Kwon, and J. C. Ye, “Diffusionclip: Text-guided diffusion models for robust image manipulation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426–2435.
-  A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, “Prompt-to-prompt image editing with cross attention control,” 2022.
-  B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, “Imagic: Text-based real image editing with diffusion models,” in *Conference on Computer Vision and Pattern Recognition 2023*, 2023.
-  O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, “Text2live: Text-driven layered image and video editing,” in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*. Springer, 2022, pp. 707–723.

Referencias II

-  N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, “Plug-and-play diffusion features for text-driven image-to-image translation,” *arXiv preprint arXiv:2211.12572*, 2022.
-  Z. Zhang, L. Han, A. Ghosh, D. Metaxas, and J. Ren, “Sine: Single image editing with text-to-image diffusion models,” 2022.
-  G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, “Zero-shot image-to-image translation,” 2023.
-  I. Han, S. Yang, T. Kwon, and J. C. Ye, “Highly personalized text embedding for image manipulation by stable diffusion,” *arXiv preprint arXiv:2303.08767*, 2023.
-  Q. Wang, B. Zhang, M. Birsak, and P. Wonka, “Mdp: A generalized framework for text-guided image editing by manipulating the diffusion path,” *arXiv preprint arXiv:2303.16765*, 2023.