Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Vision Transformers para estimación de precipitación a partir de datos satelitales

Tesis de Licenciatura en Ciencias de Datos

Tobias Muñoz

Directores: Dr. Juan José Ruiz y Dr. Pablo Augusto Negri
Codirector: Lic. Sergio Hernán González

Buenos Aires, 2024

A mi familia...

AGRADECIMIENTOS

Quiero agradecer inicialmente a Juan Ruiz por proponer un tema de tesis que me llamó fuertemente la atención y haber aceptado dirigir esta tesis. A Pablo Negri por sus recomendaciones a lo largo de la tesis, acercamiento de papers, y grandes aportes e ideas desde la perspectiva de machine learning. Y a Sergio Gonzalez por estar presente en cada una de las reuniones semanales para discutir avances.

También agradecer al Centro de Investigaciones del Mar y la Atmósfera (CIMA) por brindarme la posiblidad de usar sus servidores, que de otra forma no hubiese podido llevar a cabo de forma completa esta tesis dado el costo computacional y de almacenamiento requerido para correr el modelo y los experimentos.

A mi familia por su apoyo desde el comienzo de mi vida universitaria. Sin ellos no hubiese podido ser el primer graduado del bachiller universitario en ciencias de datos de toda la facultad, ni haberme dedicado de forma completa al estudio. Todo mi sacrificio y esfuerzo a lo largo de estos años fue posible gracias a ellos.

Resumen

Los fenómenos meteorológicos de alto impacto social como las tormentas intensas, generan daños a la población humana y cuantiosas pérdidas materiales todos los años. En el Servicio Meteorológico Nacional (SMN) se desarrollan técnicas para mejorar el pronóstico a corto plazo (horas a días) de dichos fenómenos con el objetivo final de proteger a la población y sus bienes. Entre las herramientas que se utilizan habitualmente para el pronóstico de eventos extremos que puedan tener un impacto negativo en la sociedad, se cuentan el monitoreo con sensores remotos como los radares y satélites meteorológicos y las simulaciones numéricas que permiten anticipar a futuro la evolución de la atmósfera.

En esta tesis se llevó a cabo el desarrollo de un modelo de estimación cuantitativa de la precipitación a partir de información satelital utilizando la arquitectura de Vision Transformer (ViT). El ViT se alimenta con la estimación de precipitación en superficie del producto RRQPE (Rainfall Rate and Quantitative Precipitation Estimation) provenientes del sensor ABI (Advanced Baseline Imager) a bordo del satélite geostacionario GOES-16 (Geostationary Operational Environmental Satellites). Se entrena con las tasas de precipitación instantáneas estimadas a partir del radar DPR (Dual-frequency Polarization Radar) a bordo del satelite de orbita baja GPM (Global Precipitation Measurement Mission).

Los resultados obtenidos son alentadores y se demuestra el potencial de los Vision Transformer para ser utlizados como herramienta para la estimación de tasa de precipitación en escalas de tiempo del orden de los 10 minutos. A su vez, se compara con una arquitectura U-Net, con la cual se obtienen resultados similares en cuanto a las métricas pero demostrando mejores resultados en la visualización de la estimación de precipitación.

Índice general

1. INTRODUCCIÓN

1.1. Preliminares

Es fundamental conocer con la mayor precisión posible la distribución de la precipitación en diferentes escalas espacio-temporales para anticiparse a eventos meteorológicos que puedan causar cuantiosas perdidas económicas. Los eventos extremos, como las inundaciones repentinas asociadas a precipitaciones intensas, representan un desafío crítico para la protección de la población y sus bienes. Estas situaciones pueden generar daños significativos en infraestructuras, viviendas y medios de subsistencia, afectando gravemente la economía local y la calidad de vida de las comunidades afectadas. Para mitigar estos riesgos, es crucial desarrollar tecnologías de pronóstico y monitoreo que permitan anticiparse a estos fenómenos, minimizando su impacto.

La forma mas exacta de medir las precipitaciones es in-situ a traves de pluviómetros. Sin embargo, la cobertura espacial de las redes pluviométricas resulta inhomogénea como pueden verse en la figura 1.1. Ciertas regiones no poseen la densidad espacial ni la resolución temporal necesarias para caracterizar de forma adecuada la distribución de precipitaciones en escalas temporales inferiores a la diaria.

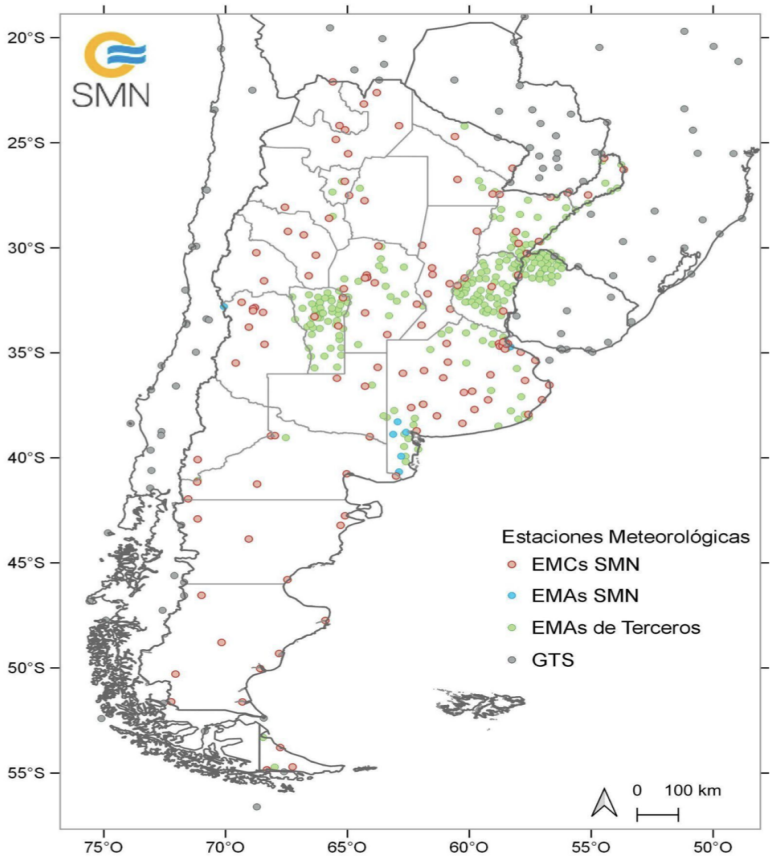

Fig. 1.1: Red de estaciones meteorológicas convencionales y automáticas del Servicio Meteorológico Nacional. Los puntos rojos corresponden a estaciones oficiales del SMN que miden cada 6 horas, los puntos verdes y azules corresponden a redes de estaciones automáticas que miden precipitacion cada 10 minutos. Los puntos violetas corresponden a redes oficiales de los paises limitrofes. Imagen extraída de Hobouchian et al. 2021.

1.2. Estimaciones de precipitación basada en sensores

Para mitigar las limitaciones presentes y complementar las redes de superficie, se desarrollaron diferentes métodos que aprovechan el potencial de sensores remotos (principalmente en radares y satélites meteorológicos) para obtener estimaciones de la tasa de precipitación. Los radares meteorológicos pueden brindar una buena estimación pero pueden haber degradaciones en su calidad producto de interferencias de señales, contaminación por ecos no meteorológicos, topografía, a la vez que su cobertura espacial es inhomogénea en las regiones continentales y prácticamente inexistente sobre los oceanos. Las estimaciones de satélites geoestacionarios en el rango del

infrarrojo presentan una gran incertidumbre debido a que la estimación de la lluvia se basa principalmente en la temperatura estimada del tope de las nubes, agregando información pero limitada para inferir la tasa de precipitación. Por otro lado, los sensores basados en microondas proveen estimaciones de precipitación más precisas que los sensores infrarrojos, ya que son sensibles al contenido de agua o hielo presentes en las nubes. Sin embargo, los sensores de microondas se encuentran a bordo de satélites de órbita baja y tienen una resolución espacial aproximadamente hasta 5 veces más baja que los de infrarrojo. Es decir, que aunque la calidad de las estimaciones es significativamente superior a las derivadas por sensores infrarrojos, su resolución espacial y temporal resulta escasa.

1.3. Estimaciones basadas en machine learning

Debido al gran avance de técnicas de machine learning y capacidad de procesamiento de datos para entrenar distintos algoritmos, se han incorporado diversas técnicas al desarrollo de estimaciones de precipitación, basados en sensores remotos. De esta forma, se puede extraer información de manera eficiente y reconociendo ciertos patrones en los datos, que a diferencia de los métodos clásicos pueden llegar a pasar desapercibidos. Uno de los primeros algoritmos en incorporar redes neuronales en la estimación de precipitación fue el algoritmo Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) (Hsu et al. 1997). Modelos posteriores de redes neuronales profundas de tipo convolucional como el PERSIANN-CNN (Sadeghi et al. 2019) y la red generativa adversarial PERSIANN-cGAN (Hayatbini et al. 2019) han demostrado un mejor desempeño en la estimación de precipitación frente a los primeros modelos de aprendizaje automático. La razón principal se debe a que estas redes profundas pueden extraer información sobre la morfología de los datos obtenidos por diferentes sensores.

En los últimos años, los Transformers (Vaswani et al. 2017) han revolucionado múltiples campos del aprendizaje profundo, extendiéndose más allá del procesamiento del lenguaje natural hacia la visión por computadora, con los Vision Transformers (ViT) como una de las principales innovaciones, donde se demostró que los Transformers pueden ser aplicados exitosamente a tareas de visión, logrando un rendimiento comparable o superior a las redes convolucionales tradicionales (Dosovitskiy et al., 2020). El artículo "SRViT: Vision Transformers for Estimating Radar Reflectivity from Satellite Observations at Scale" (Jason Stock et al. 2024) presenta una red neuronal basada en transformers diseñada para generar campos sintéticos de reflectividad de radar de alta resolución (3 km) a partir de imágenes satelitales geoestacionarias. Este enfoque tuvo como objetivo mejorar las previsiones a corto plazo de eventos meteorológicos de alto impacto y facilitar la asimilación de datos para la predicción numérica del tiempo en los Estados Unidos. Los resultados muestran una mayor nitidez y precisión en comparación con enfoques convolucionales tradicionales.

1.4. Objetivos de la tesis

La presente tesis propone desarrollar un modelo de estimación cuantitativa de precipitación basado en Vision Transformers que funcione en base a la combinación de múltiples fuentes de información, proveniente de sensores remotos. Este modelo busca contribuir al monitoreo y pronóstico de eventos meteorológicos de alto impacto social, vinculado a la ocurrencia de precipitaciones extremas. El Vision Transformer será entrenado comparando datos de la estimación de precipitación en superficie del GOES-16 con las estimaciones de precipitación obtenidas a partir del radar Dual-frequency Precipitation Radar (DPR, Iguchi et al. 2018) a bordo del satélite Global Precipitation Mission (GPM).

2. TRANSFORMERS

2.1. Transformers

Los transformers son un tipo de modelo en machine learning que han revolucionado el campo del procesamiento del lenguaje natural y entre otros más. Introducidos por Vaswani et al. en 2017, los transformers se basan en un mecanismo de atención, particularmente la "atención auto-regresiva" (self-attention), que permite al modelo enfocarse en diferentes partes de una secuencia de datos (como palabras en un texto) en paralelo de manera eficiente.

A diferencia de modelos anteriores, como las redes neuronales recurrentes (RNNs), que procesan la información de manera secuencial, los transformers permiten captar relaciones de largo alcance entre elementos de la secuencia sin necesidad de procesarla en orden. Esto los hace más rápidos y efectivos, especialmente para tareas como traducción automática, resumen de textos o, incluso, procesamiento de imágenes y series temporales, como se ve en modelos de visión por computadora y predicción de eventos.

Para procesar la información, el modelo ingresa los datos por varias capas de transformación, cada una afinando la comprensión. También usa un componente llamado codificador y decodificador: el codificador toma la entrada (como una frase), y el decodificador genera una salida (como la traducción de la frase a otro idioma).

Además, como el modelo no tiene una estructura secuencial implícita, depende de la codificación posicional (positional encoding), que agrega información de la posición de cada palabra en la secuencia a través de funciones trigonométricas, permitiendo al modelo captar las relaciones entre los tokens en una secuencia ordenada.

La arquitectura (Figura 2.1) y su explicación se observa a continuación:

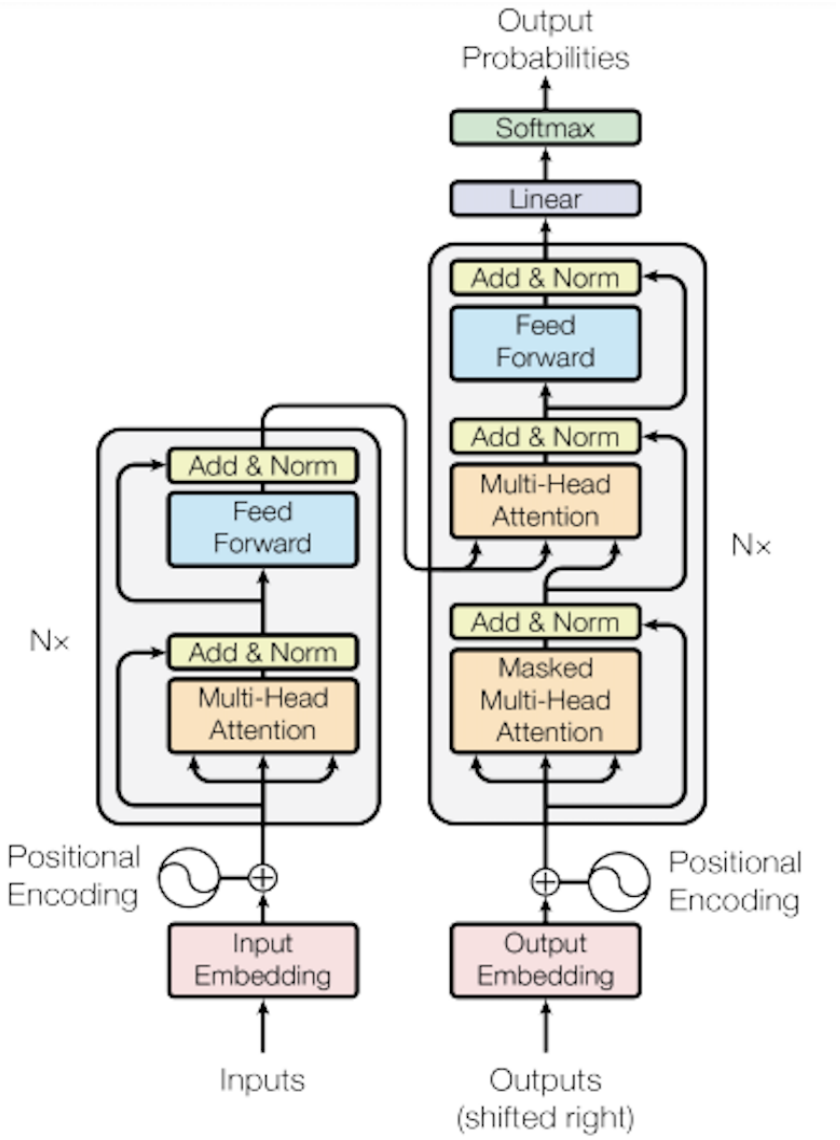

Fig. 2.1: Transformer - Arquitectura del modelo (imagen extraída de Vaswani et al. 2017)

En el codificador, cada capa incluye dos subcapas:

1. Una **capa de atención multi-cabeza** (MultiHead), que permite a cada posición en la entrada enfocarse en otras posiciones para captar relaciones entre ellas. En lugar de usar una sola operación de atención, el modelo utiliza varias **cabezas de atención** (head), que se concatenan y calculan diferentes aspectos de la relación entre los elementos de la secuencia.

Esto se expresa como:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O$$

donde W^O es la matriz de pesos que combina las salidas de todas las cabezas y cada cabeza de atención (attention) se calcula de la siguiente manera:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

Aquí, Q (queries), K (keys) y V (values) son matrices que representan las entradas de la capa. Las matrices de pesos W^Q, W^K, W^V y W^O son los parámetros que el modelo aprende durante el entrenamiento. Las dimensiones de estas matrices son de $(longitud\ de\ la\ secuencia, d_k)$, en cada cabeza de atención.
La longitud de la secuencia es el número de tokens o palabras en la entrada que el modelo procesa a la vez, mientras que d_{model} representa la dimensión total de la representación en el modelo Transformer. La operación de atención se calcula para cada cabeza como:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Por lo general, se define a d_k como $\frac{d_{model}}{h}$, donde h es el número de cabezas de atención. Este escalado por $\sqrt{d_k}$ ayuda a estabilizar el cálculo de la atención al evitar que los valores en la función softmax se vuelvan demasiado grandes. La función **softmax** convierte un vector de valores **z** en una distribución de probabilidad, donde z_i es el valor del i-ésimo elemento del vector de entrada, y la suma en el denominador se realiza sobre todos los elementos j del vector:

$$softmax(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

Esta subcapa está seguida de una **normalización de capa** (LayerNorm) y una **conexión residual**, que se expresa como:

$$LayerNorm(x + Sublayer(x))$$

La normalización de capa es una técnica que ajusta y escala los valores de entrada para mejorar la estabilidad y eficiencia del modelo. A diferencia de la normalización por lotes (Batch Normalization), que utiliza estadísticas calculadas sobre el lote de entrenamiento, LayerNorm opera independientemente en cada instancia y calcula la media y desviación estándar para cada ejemplo, lo cual es adecuado para secuencias en las que cada posición en la secuencia debe ser normalizada de forma autónoma. La conexión residual suma la entrada original **x** a la salida de **Sublayer(x)** (subcapa de Atención Multi-Cabeza) ayudando a combatir el problema del desvanecimiento del gradiente y facilitando el flujo de la información a través de las capas, especialmente en redes profundas.

2. Una **capa de red feed-forward** (FFN, por sus siglas en inglés) que aplica una transformación no lineal, procesando cada posición de forma independiente. Este proceso puede describirse matemáticamente como:

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2$$

donde W_1, W_2 son matrices de peso, b_1, b_2 son términos de sesgos y ReLU (Rectified Linear Unit) es una función de activación definida como:

$$ReLU(z) = \text{máx}(0, z)$$

permitiendo que la red aprenda relaciones no lineales entre las características, lo que es esencial para representar patrones complejos en los datos. Al igual que la subcapa de atención multi-cabeza, la red feed-forward también está seguida por normalización y una conexión residual

$$LayerNorm(x + FFN(x))$$

El decodificador de un transformer es similar al codificador, pero incluye una capa de atención cruzada que permite que la salida del codificador se combine con la secuencia generada hasta ese momento. Esta característica es fundamental para facilitar la predicción de la siguiente palabra en la secuencia de salida. La principal diferencia entre el multi-head attention y el masked multi-head attention radica en el acceso a las posiciones de la entrada: en el multi-head attention regular, el modelo puede atender a todas las posiciones de la secuencia de entrada, lo que le permite captar relaciones contextuales completas. En contraste, el masked multi-head attention, utilizado en el decodificador, aplica un enmascaramiento que impide que el modelo vea posiciones futuras en la secuencia de salida durante el entrenamiento. Esto es crucial para tareas de generación de texto, ya que garantiza que el modelo realice predicciones de manera autoregresiva, basándose únicamente en las palabras que ya se han producido. El término output (shifted right) se refiere a que la entrada al decodificador es la secuencia de salida desplazada en una posición. Por ejemplo, si la secuencia deseada es "[La, casa, es, grande]", el modelo alimenta al decodificador con "[<START>, La, casa, es]", donde <START> indica el comienzo de la secuencia. Esto permite que el decodificador genere la salida de manera autoregresiva, utilizando las palabras previamente generadas para predecir la siguiente.

2.2. Vision Transformers

Los Vision Transformers (ViT) llevan el enfoque de los transformers al campo de la visión por computadora. Los ViT dividen una imagen en pequeños parches (proceso llamado Patch embedding), tratándolos como si fueran palabras en una secuencia. Cada parche se proyecta a un espacio de características de alta dimensión (Linear Projection), conocido como la dimensión del modelo, lo que permite al modelo aprender representaciones ricas de cada fragmento de la imagen. Luego, se aplica el mismo mecanismo de auto-atención que en los transformers tradicionales, permitiendo a los ViT capturar relaciones entre diferentes regiones de la imagen de manera eficiente y eliminando la necesidad de redes convolucionales profundas.

La arquitectura de un ViT se compone de varias capas transformadoras, cada una con múltiples cabezas de atención que permiten al modelo enfocarse en diferentes partes de la imagen simultáneamente. Además, la profundidad del modelo, determinada por la cantidad de capas transformadoras, influye en su capacidad para aprender patrones complejos. En la Figura 2.2 se puede observar el tratamiento de la entrada y su posterior procesamiento en la capa transformer de forma análoga a los transformers tradicionales.

En el paper de Dosovitskiy et al., 2020, el ViT fue aplicado para tareas de clasificación, usando en la entrada un token clasificador entrenable (extra learnable embedding). Sin embargo, removiendo el token clasificador y haciendo unos ligeros cambios en la salida, como cambiar la función de activación y realizar un proceso inverso del Patch Embedding, se puede transformar sencillamente en un problema de regresión o en un problema de imagen a imagen, dejando en evidencia la versatilidad del modelo para distintas tareas.

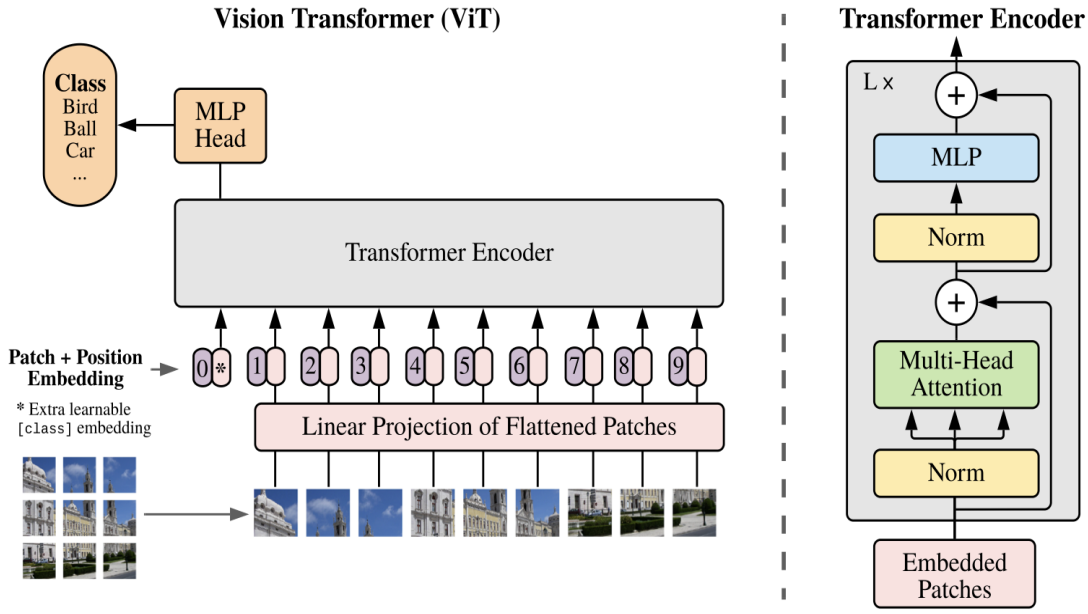

Fig. 2.2: Modelo Vision Transformer - Dosovitskiy et al. 2021

2.3. Entrenamiento

Una vez definida la arquitectura del ViT, se procede a establecer una función de costo que se optimizará durante el entrenamiento. Existen diversas funciones de costo, cada una con características específicas; sin embargo, el error cuadrático medio (MSE, por sus siglas en inglés) es uno de los más utilizados. A pesar de su popularidad, el MSE puede no ser efectivo en situaciones donde los datos presentan un desbalance significativo. Para abordar este problema, se introduce el MSE ponderado (Weighted MSE), que se define y calcula de la siguiente manera:

$$WeightedMSE = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot (y_i - \hat{y}_i)^2$$

Donde N es el número total de muestras, w_i es el peso asignado, y_i es el valor verdadero, y $\hat{y}_i$ es el valor predicho por el modelo para la i-ésima muestra. Esta fórmula permite asignar un peso a los valores

que son menos frecuentes, lo que puede ser crucial según la naturaleza del problema que se desea resolver. Los pesos se asignan a cada error cuadrado según la importancia relativa de cada observación o clase dentro del contexto del problema. Estos pesos pueden obtenerse de diferentes maneras. Una opción es definirlos en función de la relevancia de ciertos errores, asignando más peso a aquellos que representan situaciones críticas o donde los errores tienen un mayor impacto. También es común basar los pesos en la probabilidad o frecuencia de aparición de ciertos valores, asignando mayor peso a aquellos menos frecuentes para que el modelo preste más atención a ellos

Además, existe otra función de costo altenativa para este proposito conocida como quantile loss, que se utiliza especialmente en problemas de regresión donde se busca predecir intervalos de confianza en lugar de valores puntuales. Esta función de pérdida es útil para modelar situaciones en las que se desea capturar la variabilidad de los datos y no solo su tendencia central. Al minimizar la quantile loss, el modelo puede ajustarse para prever diferentes cuantiles (q) de la distribución de los datos, lo que resulta en predicciones más robustas e informativas. Su expresión viene dada por:

$$QuantileLoss(y_i, \hat{y}_i, q) = \frac{1}{N} \sum_{i=1}^{N} (q \cdot \text{máx}(0, y_i - \hat{y}_i) + (1 - q) \cdot \text{máx}(0, \hat{y}_i - y_i))$$

En este contexto, la elección del optimizador se vuelve crucial, ya que influye significativamente en la convergencia y efectividad del modelo. Los optimizadores son algoritmos que ajustan los pesos del modelo durante el entrenamiento, minimizando la función de pérdida. Entre los optimizadores más utilizados se encuentran Stochastic Gradient Descent (SGD), que actualiza los pesos utilizando un pequeño subconjunto de datos (batch), lo que lo hace eficiente en términos de memoria, y cuya expresión se define como:

$$\theta_{t+1} = \theta_t - \alpha \cdot \nabla J(\theta_t)$$

donde θ_t son los parámetros del modelo en el tiempo t, α es la tasa

de aprendizaje, y $\nabla J(\theta_t)$ representa el gradiente de la función de costo J calculada sobre el mini-batch.

Otro optimizador común es Adam, que combina las ventajas de SGD con el uso de momentos, permitiendo un ajuste adaptativo de las tasas de aprendizaje. La actualización de los parámetros en Adam se define como:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t$$

donde $\hat{m}_t$ es la estimación del primer momento (media), $\hat{v}_t$ es la estimación del segundo momento (varianza), y ϵ es un término pequeño para prevenir divisiones por cero. En este contexto, $\hat{m}_t$ puede interpretarse como la **velocidad** de los cambios en los parámetros del modelo, ya que indica la dirección y magnitud del ajuste, mientras que $\hat{v}_t$ se asemeja a la **aceleración**, ya que mide la variabilidad de los gradientes y ayuda a adaptar la tasa de aprendizaje de manera más efectiva.

3. BASE DE DATOS

3.1. Datos

El Vision Transformer que se desarrolla en este trabajo busca estimar la distribución espacial de la precipitación, tomando como entrada las estimaciones de las tasas de precipitación en superficie (RRQPE) provenientes del sensor ABI a bordo del satélite GOES-16, y del radar DPR a bordo del satélite GPM para las tasas de precipitación, con el objetivo de generar un mapa preciso de las precipitaciones a lo largo de grandes áreas geográficas.

El GOES-16 proporciona datos de estimación de precipitación en el infrarrojo mediante varios canales, en especial el canal 13 (10,3 μm) del sensor ABI (ver figura 3.1). Este satélite de órbita geoestacionaria (centrado en la latitud 0^{o} y longitud $75{,}2^{o}$ O), cubre casi la totalidad del continente americano y los océanos adyacentes, ofreciendo imágenes desde febrero de 2018 para las estimaciones de precipitacion basadas en los datos provistos por el ABI (algoritmo RRQPE). El canal 13 es útil por su capacidad de medir la temperatura de los topes nubosos o del suelo (en cielos despejados), con una resolución espacial de 2 km (en el punto nadir, estando debajo del satélite) y una frecuencia temporal de 10 minutos.

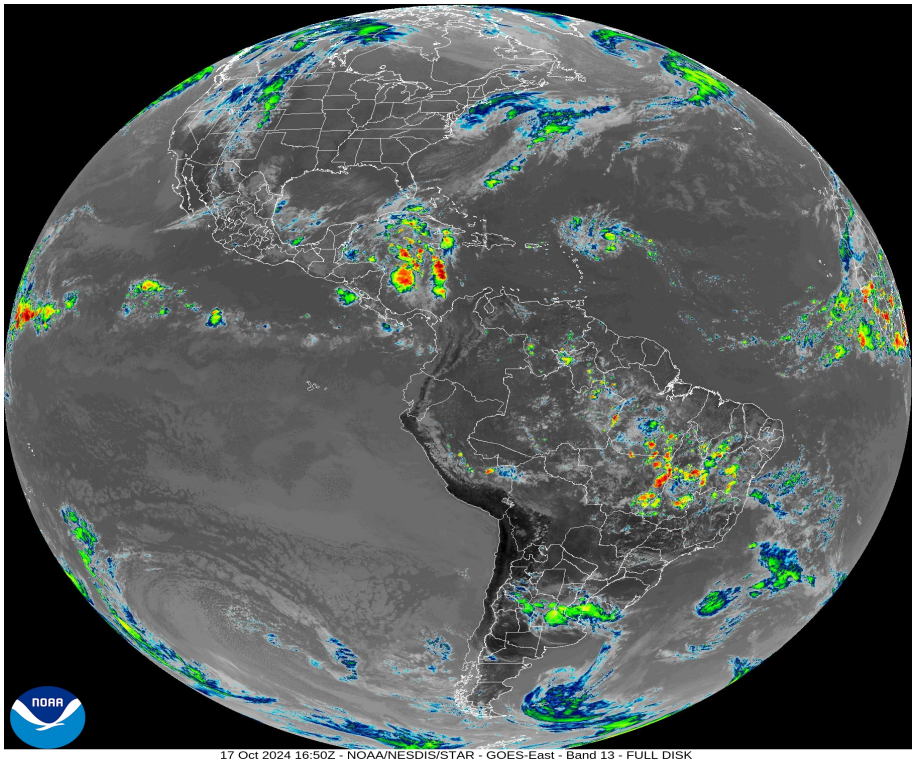

Fig. 3.1: Imagen de escaneo de disco completo del canal 13 del sensor ABI a bordo del satélite GOES-16. Los colores representan diferentes temperaturas de la superficie de las nubes y el suelo, donde los colores frios (azul, verde) indican temperaturas más bajas, y colores mas calidos (amarillo, naranja y rojo), suelen representar temperaturas más altas

Por otro lado, el satélite GPM aporta datos de tasas de precipitación estimadas a partir del radar de doble frecuencia Dual-frequency Polarization Radar (DPR, Iguchi et al., 2018) (ver figura 3.2), disponible desde 2014. Este radar utiliza dos frecuencias Ka (35.5 GHz) y Ku (13.6 GHz) para medir la reflectividad de las nubes, proporcionando estimaciones precisas de la precipitación. La resolución temporal es de 1 a 2 días. El sensor DPR cuenta con resolución espacial de 5.2 km en la horizontal para ambas bandas. En cuanto a la sensibilidad de estimación, la banda Ku presenta un mínimo de detección de 0.5 mm/h, y la banda Ka una detección mínima de 0.25 mm/h, siendo ésta última de alta sensibilidad proveyendo mejor información asociada a precipitaciones débiles.

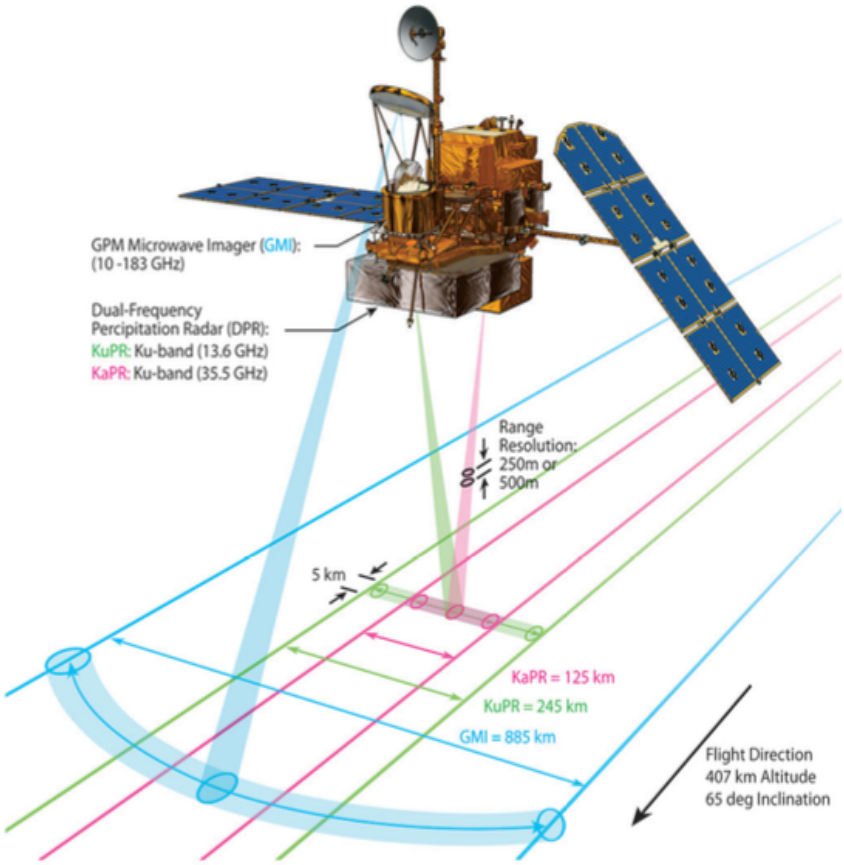

Fig. 3.2: Esquema de escaneo de los sensores GMI (celeste) y DPR (banda Ka en lila y banda Ku en verde) del satélite GPM (Iguchi et al., 2018).

3.2. Región de estudio y generación de la base de datos

Como región de estudio se eligió a Sudamérica y sus oceanos adyacentes entre las latitudes 50º S y 10º N, y longitudes 90º O y 30º O, incluyendo así casi la totalidad del territorio argentino. Es fundamental la elección de la región de estudio para el entrenamiento de la arquitectura dado que una región muy pequeña puede facilitar el entrenamiento pero reduce la cantidad de información disponible. Por otro lado, una región muy amplia puede incrementar la complejidad del modelo para representar mayores situaciones debido a la heterógeneidad de los valores de precipitación a lo largo de la región.

Como lo que se quiere estimar son las tasas de precipitación instantáneas, a partir de un conjunto de imágenes satelitales en el rango del infrarrojo, se construye una base de datos para el entrenamiento del modelo. Esta base consiste en 2 conjuntos de datos: uno proveniente del sensor ABI. El otro conjunto consiste en las estimaciones de precipitación obtenidas por medio del radar DPR a bordo del satélite GPM.

Para el entrenamiento se seleccionó el periodo comprendido entre el 1 de Enero de 2020 y el 30 de Junio de 2021 (1 año y medio) en sectores en donde se haya podido obtener de forma simultánea la información del satélite GOES-16, y la estimación de lluvia obtenida a partir del satélite GPM. Se obtuvieron así, sectores de 240 km de ancho y 240 km en la dirección paralela al desplazamiento del satélite. La resolución resultante de las imagenes es de 48x48 píxeles unificadas a una resolución de 5 km, donde la diferencia temporal entre la imagen del DPR y la imagen del ABI no supera los 5 minutos de forma absoluta.

La figura 3.3 muestra la región tomada para cada satélite, las estimaciones de precipitación para el GPM y la temperatura de brillo para el GOES (datos usados por el RRQPE para la estimación de precipitación).

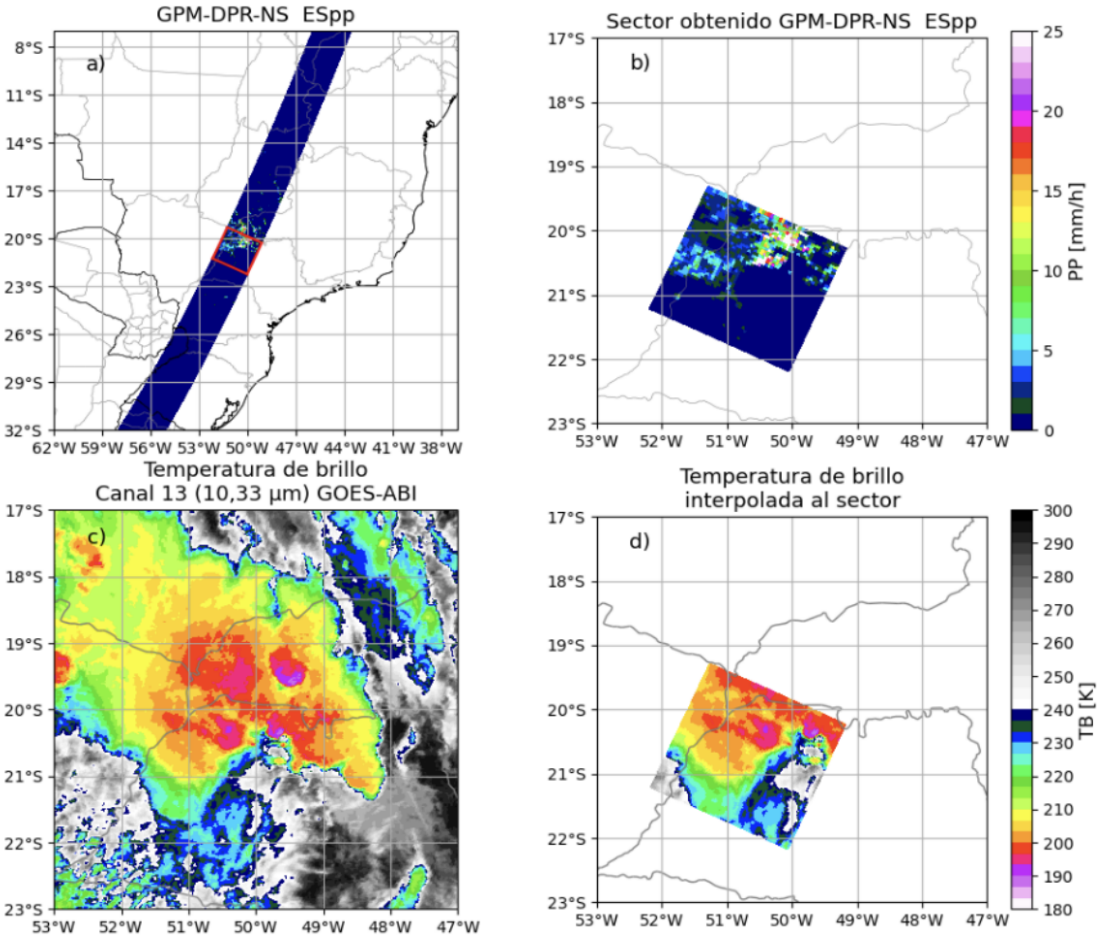

Fig. 3.3: a) Estimación de precipitación para la variable correspondiente al DPR del satélite GPM. b) Sector obtenido de a) (recuadro rojo). c) Imagen del canal 13 del GOES-16 más cercana en tiempo al sector mostrado en b). d) Imagen interpolada de c) a partir de la geometría de b).

4. METODOLOGÍA

4.1. Conformación del conjunto de datos de entrenamiento, validación y testeo

Previo al entrenamiento del modelo, se realiza una selección de imágenes donde tengan al menos un 15 por ciento de tasas de precipitación mayores a 0.1 mm/h y se eliminaron aquellos datos que poseen precipitación mayor a 275 mm/h dado que resultan ser datos erroneos (la eliminacion de estos datos es a nivel de pixel). El conjunto final incluye aproximadamente 26.500 imágenes, divididas en subconjuntos para entrenamiento, validación y testeo. La cantidad resultante para cada grupo resulta ser de aproximadamente 13.000, 1.500 y 12.000, respectivamente.

Idealmente, se busca que la distribución de probabilidad de precipitación estimada sea similar en estos 3 subconjuntos. La figura 4.1 muestra la frecuencia de las tasas de precipitación sobre el conjunto de entrenamiento, validación y testeo. Puede verse que la forma y patrón de cada histograma es muy similar para cada uno de los subconjuntos. Los datos de testeo tomados son aquellos en el periodo de enero y junio (un mes correspondiente al verano y otro al invierno del Hemisferio Sur). Se puede observar que en los datos de testeo se observa un pico en precipitaciones entre 275 y 300 mm/h, debido a que a los datos de testeo no se les aplicó ningún filtro.

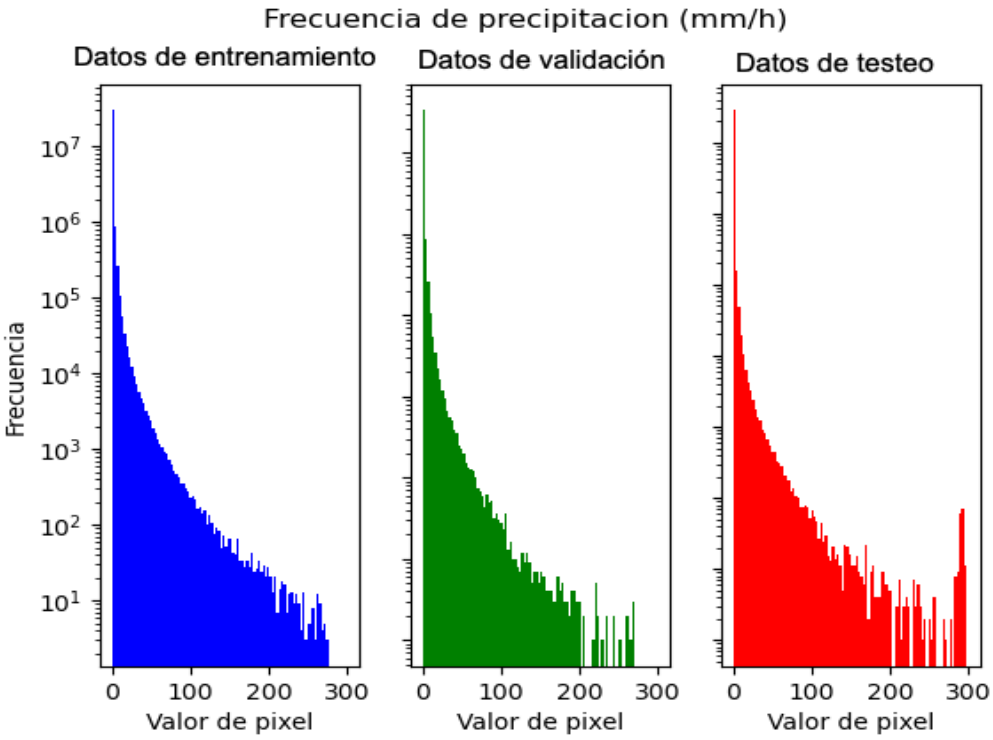

Fig. 4.1: Frecuencia de precipitación para los subconjuntos de entrenamiento, validación y testeo

4.2. Arquitectura del modelo

La arquitectura del modelo Vision Transformer empleada en este trabajo es encoder-only dado que para tareas de imagen a imagen resulta apropiado cuando la salida tiene la misma estructura espacial que la entrada. Esto permite reducir la complejidad del modelo manteniendo la capacidad de capturar relaciones espaciales complejas.

Se utiliza un tamaño de parches de 4x4 con solapamiento del 50 por ciento, lo cual permite capturar información espacial relevante para la predicción de precipitación. La dimensión del modelo es 16, lo que significa que cada parche de la imagen es proyectado a un vector de 16 dimensiones. Cada transformer cuenta con 4 cabezas de atención para el mecanismo de autoatención, lo que permite al modelo procesar múltiples relaciones espaciales a diferentes escalas, y 3 capas de transformadores, que equilibran la profundidad del modelo sin

sobreentrenar. El perceptron multicapa (MLP, por sus siglas en inglés) posee dos capas lineales con una capa oculta (16 neuronas para las externas y 64 para la oculta) y activación ReLU, lo que es crucial para aprender patrones complejos en los datos.

La decisión de tomar un tamaño de parches de 4x4 con solapamiento del 50 por ciento es debido a que intuitivamente al ser un tamaño considerablemente pequeño respecto a la imagen original (48x48) esto podría estimar de forma más precisa la estimación de precipitación a nivel pixel. En cuanto a la dimensión del modelo se escogió 16 dado que la proyección lineal de los parches de 4x4 (16 pixeles en total) supone un mapeo de forma más natural. El modelo cuenta con 4 cabezas de atención dado que se requiere que el cociente entre la dimensión del modelo y cantidad de cabezas sea un número entero, y 3 capas de transformadores por simplicidad.

Finalmente, la salida del MLP se pasa a la función inversa del patch embedding para reconstruir la imagen de salida en la misma resolución espacial que la entrada.

El proceso de optimización de los pesos del modelo durante el entrenamiento se realiza utilizando el algoritmo ADAM con una tasa de aprendizaje (learning rate) inicial de 0.001 adaptable según el rendimiento en la función de costo durante el entrenamiento. Se tomó un batch size de 16, y se entrenó durante 5 épocas. La función de costo empleada es el MSE ponderado, donde los pesos se calcularon como los valores de los datos a estimar divididos por el máximo del lote (batch). Adicionalmente se agregaron unos pesos para poner énfasis en distintos umbrales, asignando un peso distinto a umbrales bajos (mm/h < 2), medios ($2 <$ mm/h < 15) y altos (mm/h > 15)

Respecto a la arquitectura y proceso de optimización de la red U-Net se puede encontrar información detallada en el trabajo de González et al. 2024.

4.3. Métricas de evaluación

Para evaluar el desempeño del modelo, se utilizan dos métricas: en primer lugar, se utiliza la raíz del error cuadrático medio (RMSE), que mide la magnitud del error promedio entre los valores estimados por el modelo y los valores de referencia. Y por último, el coeficiente de correlación (r), que indica en que grado se relacionan linealmente los conjuntos de datos, proporcionando una medida de la fuerza y dirección de la relación entre ambos.

Asimismo, se incorporan tres métricas adicionales para evaluar la habilidad del modelo en la detección de eventos específicos dentro del rango de predicción:

- **POD (Probability of Detection)**, que mide la capacidad del modelo para identificar correctamente los eventos positivos, con un valor ideal de 1 (100 %), indicando que todos los eventos positivos fueron detectados correctamente.

$$POD = \frac{TP}{TP + FN}$$

  donde TP son los verdaderos positivos y FN los falsos negativos.

- **FAR (False Alarm Rate)**, que evalúa la proporción de predicciones incorrectas en relación a todos los positivos, donde un valor ideal es 0 (0 %), lo que significa que no hay falsas alarmas.

$$FAR = \frac{FP}{TP + FP}$$

  donde FP son los falsos positivos.

- **ETS (Equitable Threat Score)**, que proporciona una medida global del desempeño, balanceando tanto aciertos como falsos positivos y falsos negativos de manera equitativa, con un valor ideal también de 1 (100 %), lo que indica un desempeño perfecto en la identificación de eventos.

$$ETS = \frac{TP - \frac{FP \cdot FN}{N}}{TP + FP + FN - \frac{FP \cdot FN}{N}}$$

5. RESULTADOS

En esta sección se discuten los resultados obtenidos de los experimentos con el Vision Transformer. Se observó que los cambios realizados en la arquitectura no variaron significativamente los resultados, incluso cuando se incrementó la complejidad del modelo. Sin embargo, se evidenció una ligera sensibilidad en relación al tamaño de los parches utilizados. Se entrenó un modelo ViT, y se compararon los resultados contra una arquitectura U-Net (ambos entrenados con la función de costo MSE ponderada), para evaluar las diferencias en el enfoque de ambas arquitecturas y su influencia en la habilidad de modelar patrones complejos y eventos críticos en los datos.

5.1. Sensibilidad a la función de costo

La sensibilidad de la MSE ponderada es crucial para capturar de forma precisa los valores de precipitación en un rango amplio, especialmente en contextos donde los valores altos son menos frecuentes pero significativos. Al asignar mayores pesos a los errores asociados con valores de precipitación elevados, la MSE ponderada permite que el modelo priorice la minimización de estos errores. En base a este enfoque el modelo resultante es capaz de captar tasas de precipitación de hasta 25 mm/h, aproximadamente. Sin embargo, la efectividad disminuye cuando se acerca a este umbral, al ser valores muy altos y poco frecuentes. En consecuencia, el uso de la MSE ponderada requiere un balance cuidadoso entre capturar los valores bajos y moderados de precipitación, manteniendo al mismo tiempo una sensibilidad suficiente para reflejar los valores extremos. Aunque los datos abarcan precipitaciones de hasta 275 mm/h, su baja frecuencia hace que el modelo esté sesgado hacia valores pequeños, manteniendo un error bajo en estos rangos más comunes.

El efecto del desbalance de los datos sobre la prediccion se refleja en la figura 5.1, donde se observan diferencias en la distribución de las precipitaciones estimadas por el GPM y por el modelo basado en ViT. Se puede notar el sesgo que posee el modelo para estimar precipitaciones en el rango de 0 a 25 mm/h aproximadamente, y la incapacidad para estimar umbrales mas extremos.

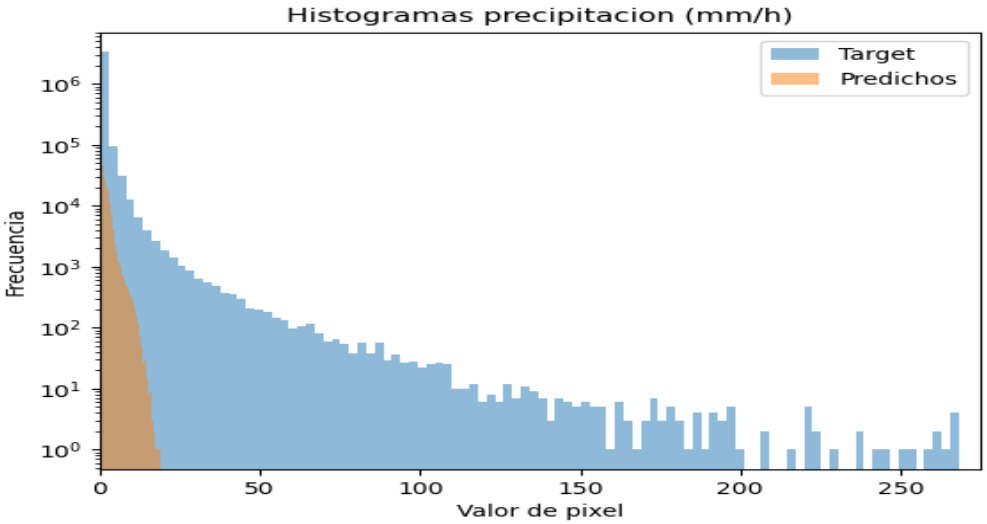

Fig. 5.1: Frecuencia de precipitación - Modelo ViT

Para las métricas FAR, POD y ETS se utilizan umbrales de 0.1, 2.5, 5, 10, 25, 50, 75 y 100 mm/h para analizar la efectividad de las predicciones (ver Figura 5.2). Para umbrales bajos, el POD y ETS muestran valores ligeramente buenos, pero a medida que el umbral se incrementa se puede ver como las métricas tienden a empeorar. Respecto a la FAR, se observa una pequeña mejora en el paso del umbral 0.1 al umbral 2.5, luego sube hasta valores cercanos a 1 y decae marcadamente a 0 para el umbral 25 en adelante, lo cual tiene sentido pues el modelo no tiene la capacidad de detectar valores mayores a 25 mm/h.

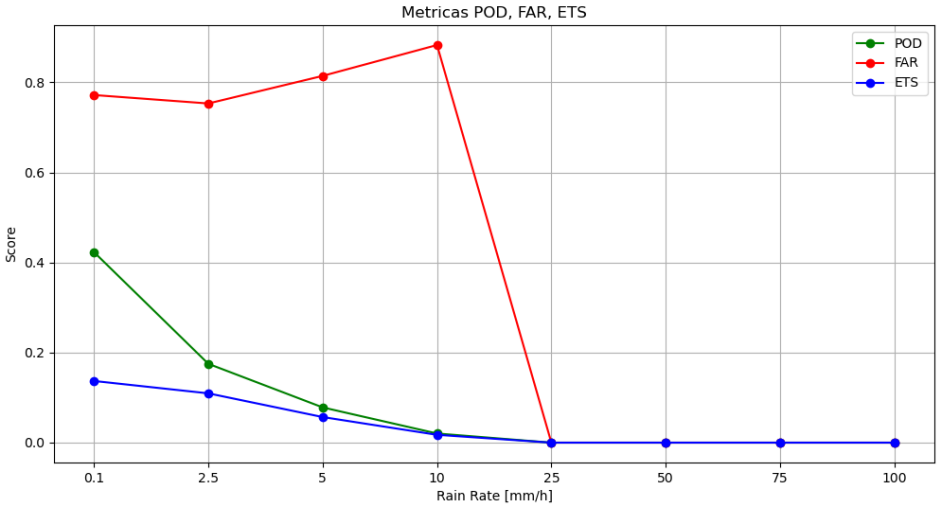

Fig. 5.2: Métricas POD, FAR, ETS - Modelo ViT

Comparando el Vision Transformer con el modelo U-Net respecto a la frecuencia de precipitación estimada se observa un patrón muy similar casi indistinguible del modelo Vision Transformer, con un umbral máximo de precipitación de 25 mm/h aproximadamente.

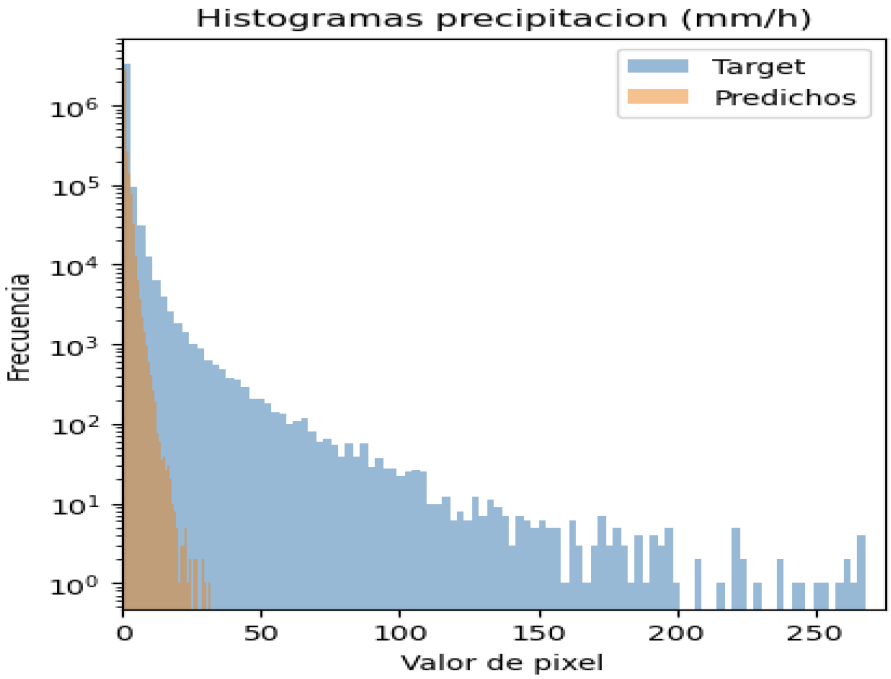

Fig. 5.3: Frecuencia de precipitación - Modelo U-Net

En cuanto a las métricas FAR, POD y ETS se observan patrones y valores similares al modelo ViT pero con diferencias significativas en el umbral 0.1 para la POD, donde el modelo Unet posee un score de 1, pero luego decae fuertemente al incrementar el umbral, la ETS que comienza en 0, y la FAR que posee score 1 para el umbral 25. En definitiva, las tres métricas resultan similares en valores y patrón.

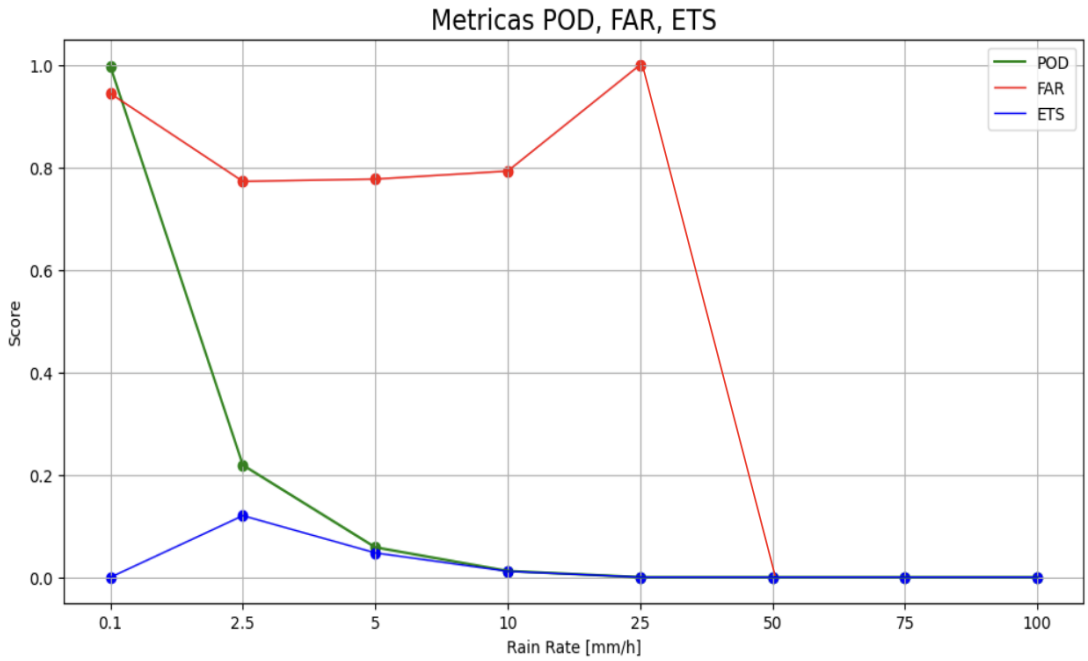

Fig. 5.4: Métricas POD, FAR, ETS - Modelo U-Net

En la Tabla 1 se presentan los resultados comparativos de las métricas RMSE y coeficiente de correlación entre las arquitecturas ViT y U-Net. A pesar de la gran variabilidad de los datos, se observa que ambos modelos presentan un RMSE que indica una buena capacidad general de estimación. Por otro lado, el coeficiente de correlación sugiere una relación positiva débil entre las predicciones y los valores reales.

Tab. 1: RMSE y Coeficiente de Correlación

Métrica	**ViT**	**U-Net**
RMSE	1.5 mm/h	1.55 mm/h
Coeficiente de correlación	0.22	0.22

Se pudo observar que en la comparación por distribución de precipitación y las métricas tanto de rendimiento del modelo para umbrales como rendimiento general, hubo leves diferencias. Sin embargo, al graficar la entrada, la salida real y la salida del modelo es donde se observan diferencias considerables y signficicativas (ver figuras 5.5, 5.6 y 5.7, 5.8). Para el caso del Vision Transformer la salida es mas suave, menos ruidosa y predice de forma correcta aquellos pixeles con precipitación nula. El modelo U-Net tiende a sobreestimar los casos de precipitación nula, prediciendo en la mayoria de los casos valores en el rango 0.1 a 0.5 mm/h.

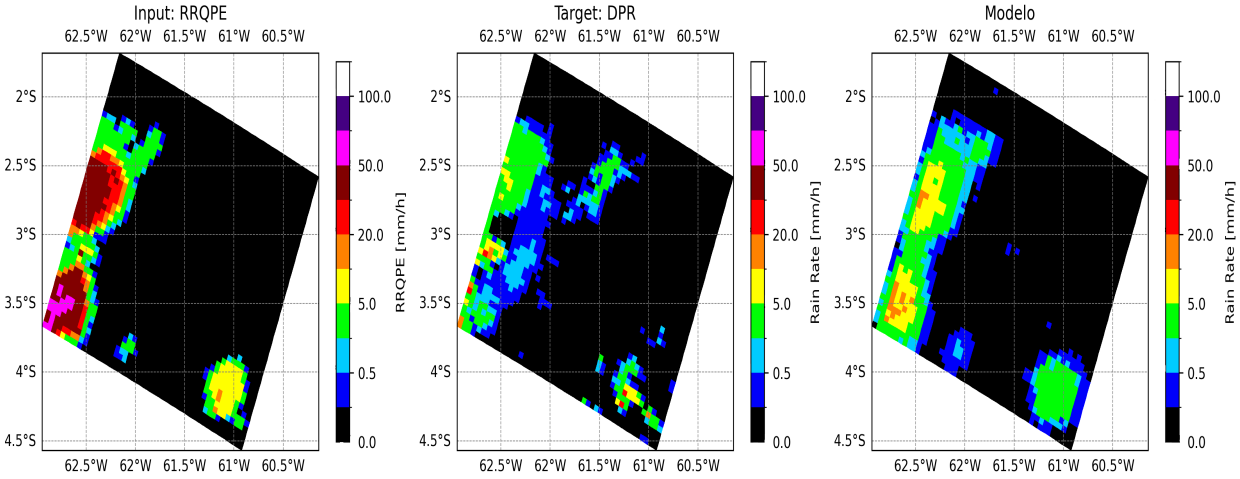

Fig. 5.5: Estimación de precipitación Caso 1 - Modelo ViT

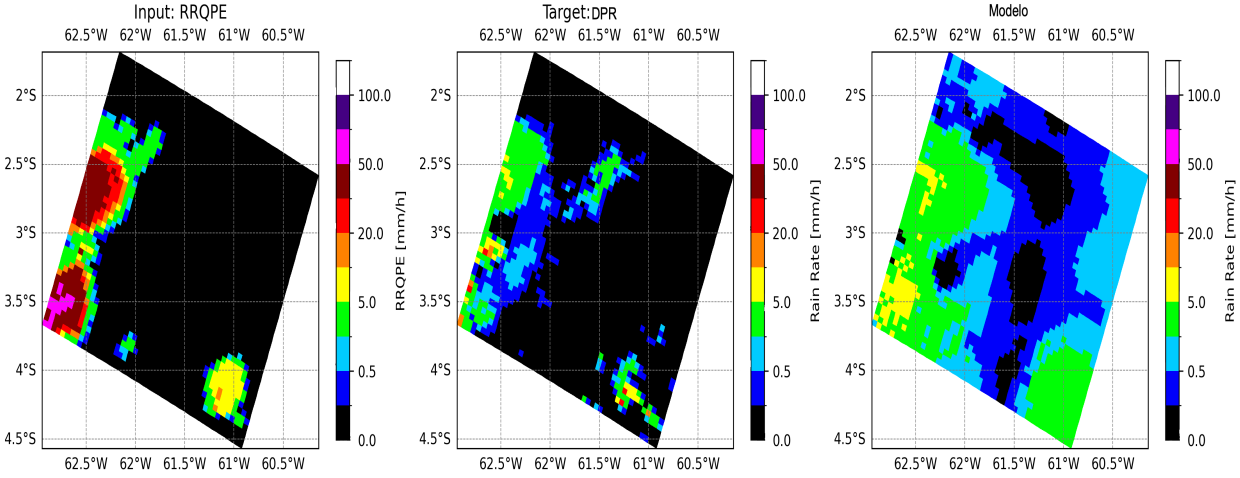

Fig. 5.6: Estimación de precipitación Caso 1 - Modelo U-Net

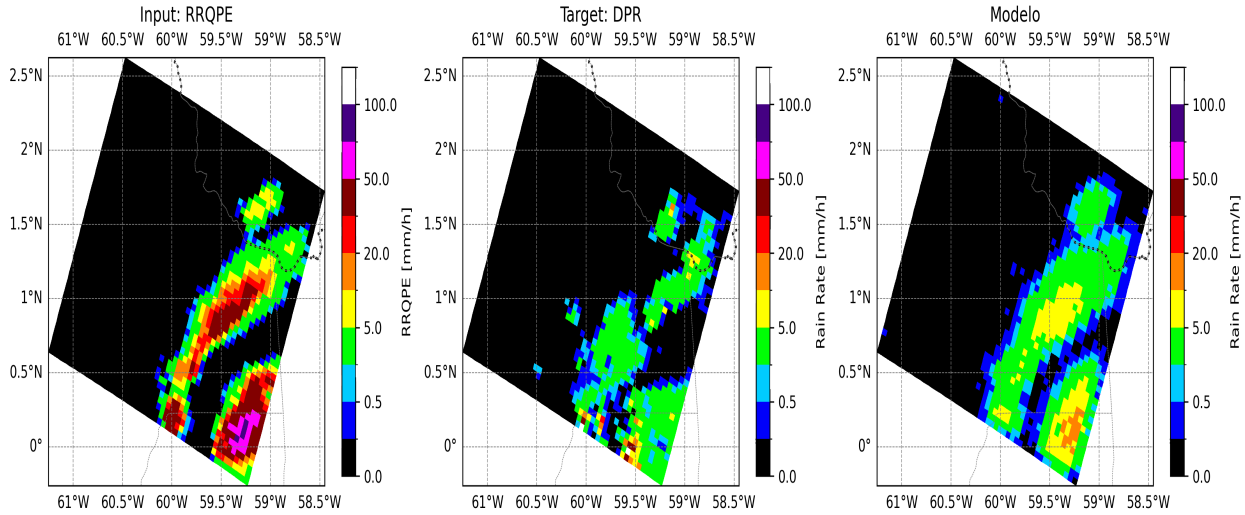

Fig. 5.7: Estimación de precipitación Caso 2 - Modelo ViT

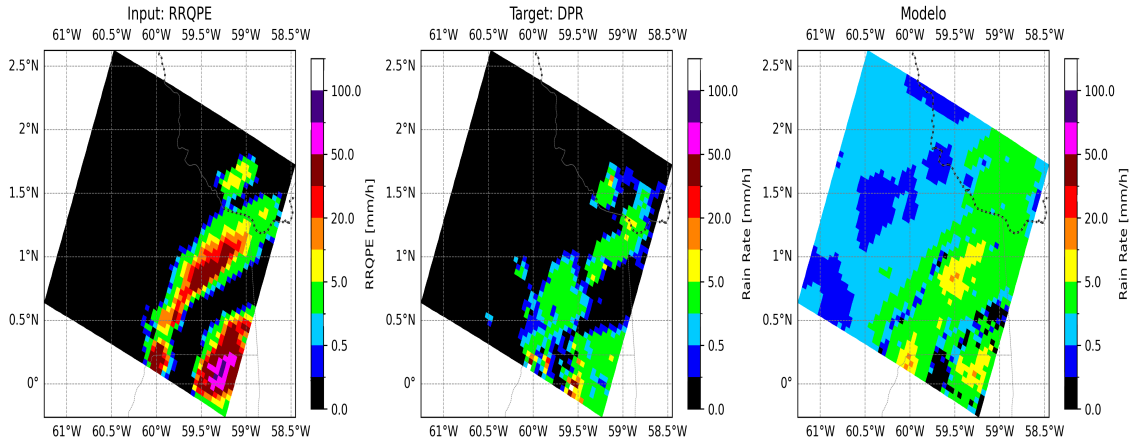

Fig. 5.8: Estimación de precipitación Caso 2 - Modelo U-Net

Además se realizaron diferentes experimentos analizando la sensibilidad de hiperparámetros en los modelos de Vision Transformer pero no se observaron cambios significativos en los resultados al ajustar parámetros como la cantidad de capas, cantidad de cabezas de atención o la dimensión del modelo. Aunque se notaron ligeras mejoras al aumentar la complejidad de la arquitectura, la simplicidad del modelo demostró ser suficiente para captar los distintos patrones en los datos. Fue en los pesos de la función de costo MSE ponderada donde se observó la única sensibilidad significativa al variar los pesos en umbrales bajos (mm/h < 2), medios ($2 <$ mm/h < 15) y altos (mm/h > 15). A su vez, el

tamaño de los parches no tuvo gran impacto en el rendimiento general del modelo, si no en la suavidad espacial general al graficar las estimaciones de precipitación para diversos casos. También se realizaron experimentos con la función de costo Quantile Loss, en los cuales se observó la capacidad del modelo para estimar valores de precipitación más extremos (hasta los 100 mm/h, aproximadamente). Sin embargo, esto fue en detrimento de las métricas utilizadas con sobrestimaciones de precipitación en la gran mayoria de los casos.

6. CONCLUSIONES

6.1. Conclusiones

El análisis del rendimiento del Vision Transformer (ViT) frente a U-Net destaca la ventaja del ViT en cuanto a simplicidad de modelo en el entrenamiento. A pesar de cambios experimentales en la arquitectura y complejidad del ViT, los resultados no presentaron variaciones significativas, mostrando que el modelo mantiene un rendimiento competitivo sin necesidad de ajustes complejos o de aumentar considerablemente las capas o parámetros. La sensibilidad del ViT al tamaño de los parches fue moderada, observándose una mejora en la suavidad de las predicciones visualizadas, lo cual favorece la reducción de ruido en las estimaciones sin requerir configuraciones adicionales.

Un aspecto clave de este análisis es el bajo número de épocas requerido por el ViT para alcanzar una precisión razonable, particularmente en comparación con U-Net. Dado que U-Net necesita entrenarse durante más épocas para reducir el ruido y hacer predicciones más suaves, el ViT representa una alternativa más eficiente, ya que logra resultados satisfactorios con un entrenamiento menos intensivo.

Además, ambos modelos presentan patrones de predicción similares en cuanto a la distribución de precipitaciones, con umbrales máximos en torno a los 25 mm/h. Sin embargo, el ViT sobresale en la suavidad y precisión en la predicción de valores nulos de precipitación, sin sobreestimarlos como ocurre en el caso del U-Net. El ViT tiene un mejor control sobre los valores mínimos, siendo menos propenso a predecir precipitaciones falsas en valores cercanos a cero, lo cual sugiere que es más confiable para estimar áreas sin precipitación.

Por otro lado, los modelos se entrenaron usando una función de costo MSE ponderada, que prioriza errores en valores altos de

precipitación. Esto permitió al ViT mantener una precisión razonable en la mayoría de los rangos de precipitación, aunque se evidenció una limitación en valores extremadamente altos debido a su baja frecuencia en los datos. No obstante, el ViT se adapta mejor a esta ponderación, balanceando la sensibilidad a valores altos sin afectar la precisión en valores comunes, lo cual es crucial en contextos de precipitación moderada.

En conclusión, el Vision Transformer demuestra una capacidad robusta para modelar patrones de precipitación con un modelo relativamente simple y con un número de épocas relativamente bajo.

Bibliografía

Hobouchian, M. P., Díaz, G. M., Vidal, L., García Skabar, Y., Ferreira, L. J., Maas, M., Rossi Lopardo, M. S., Veiga, H., & Rugna, M. (2021): *Ajuste de la estimación de precipitación satelital IMERG con observaciones pluviométricas en Argentina*. 20.500.12160/1694

Nguyen, P., Ombadi, M., Sorooshian, S., Hsu, K., AghaKouchak, A., Braithwaite, D., Ashouri, H., & Thorstensen, A. R. (1997): *The PERSIANN family of global satellite precipitation data: a review and evaluation of products*. 10.5194/hess-22-5801-2018

Sadeghi, M., Akbari Asanjan, A., Faridzad, M., Nguyen, P., Hsu, K., Sorooshian, S., & Braithwaite, D. (2019): *PERSIANN-CNN: Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks–Convolutional Neural Networks*. 10.1175/jhm-d-19-0110.1

Hayatbini, N., Kong, B., Hsu, K., Nguyen, P., Sorooshian, S., Stephens, G., Fowlkes, C., Nemani, R., & Ganguly, S. (2019): *Conditional Generative Adversarial Networks (cGANs) for Near Real-Time Precipitation Estimation from Multispectral GOES-16 Satellite Imageries—PERSIANN-cGAN*. 10.3390/rs11192193

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017): *Attention Is All You Need*. 1706.03762

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020): *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2010.11929

Stock, J., Hilburn, K., Ebert-Uphoff, I., & Anderson, C. (2024): *SRViT: Vision Transformers for Estimating Radar Reflectivity from Satellite Observations at Scale*. 2406.16955

González, Sergio & Negri, Pablo & Vidal, Luciano & Ruiz, Juan & Silvarrey, Alejo (2024), Improving instantaneous satellite rain rates estimations with machine learning. Climate informatics. 10.13140/RG.2.2.28560.83202.