Súper Resolución con Autoencoders Convolucionales

Armando Rios-Lastiri

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México

Resumen La resolución es una característica importante para determinar la naturaleza y las características de la imagen. Mejorar la resolución fortalece las características ocultas dentro de la imagen y hace que la imagen sea más nítida e informativa. Los autoencoder se han empleado anteriormente en este tipo de tareas. En este trabajo se presenta una arquitectura de acutoencoder con conexiones residuales para la restauración de imágenes de baja resolución.

1. Introducción

El objetivo central de la superresolución (SR) es generar una imagen de mayor resolución a partir de imágenes de menor resolución. La imagen de alta resolución ofrece una alta densidad de píxeles y, por lo tanto, más detalles sobre la escena original. La necesidad de alta resolución es común en las aplicaciones de visión por computadora para un mejor desempeño en el reconocimiento de patrones y análisis de imágenes. La alta resolución es de importancia en las imágenes médicas para el diagnóstico. Muchas aplicaciones requieren hacer zoom en un área específica de interés en la imagen en la que la alta resolución se vuelve esencial, p. Ej. aplicaciones de vigilancia, forense y de imágenes por satélite.

2. Contexto del problema

La súper resolución (SR) se basa en la idea de que se puede utilizar una combinación de secuencias de imágenes de una escena de baja resolución (ruidosas) para generar una imagen de alta resolución o una secuencia de imágenes. Por lo tanto, intenta reconstruir la imagen de la escena original con alta resolución dado un conjunto de imágenes observadas a menor resolución. El enfoque general considera las imágenes de baja resolución como resultado del remuestreo de una imagen de alta resolución. Entonces, el objetivo es recuperar la imagen de alta resolución que, cuando se muestrea nuevamente en función de las imágenes de entrada y el modelo de imagen, producirá las imágenes observadas de baja resolución.

3. Trabajo relacionado

Los métodos de SR basados en el aprendizaje profundo han mostrado un alto potencial en el campo de la interpolación y restauración de imágenes, en comparación con los algoritmos convencionales de interpolación por píxeles. Dong y col. propuso una estructura CNN de tres capas llamada red neuronal convolucional de superresolución (SR-CNN) — 9 —, que aprende un mapeo de extremo a extremo de una imagen de baja resolución interpolada bicúbica a una imagen de alta resolución. Desde la llegada de SR-CNN, se han desarrollado una variedad de redes CNN con una estructura de red más profunda y densa [10-13] para mejorar la precisión de SR. En particular, He et al. propuso una ResNet [11] para la clasificación de imágenes. Su idea clave es aprender los residuos a través de la conexión de salto global o local. Señala que ResNet puede proporcionar un proceso de entrenamiento de alta velocidad y evitar los efectos de desaparición del gradiente. Además de ResNet, Huang et al. propuso redes convolucionales densamente conectadas (DenseNet) [12] para combinar mapas de características jerárquicas disponibles a lo largo de la profundidad de la red para representaciones de características más flexibles y ricas. Dong y col. propuso una CNN de reducción de artefactos (AR-CNN) [14], que reduce eficazmente los diversos artefactos de compresión, como los artefactos de bloque y los artefactos de timbre en las imágenes de compresión del Joint Photographic Experts Group (JPEG).

4. Marco teorico

Single image super resolution (SISR) se refiere a la reconstrucción de una imagen de alta resolucion (HR) a partir de una observación de baja resolucion (LR). Dada una imagen LR $\mathbf{Y}$, generalmente se supone que está degradada a partir de una imagen HR correspondiente X, que se puede representar como

$$\mathbf{Y} = D\left(\mathbf{X}, \theta_D\right)$$

donde $D(\cdot)$ denota el proceso de degradación definido por el conjunto de parámetros θ_D. En un escenario real, el parámetro de degradación θ_D es desconocido, y todo lo que tenemos es la imagen LR $\mathbf{Y}$. SISR tiene como objetivo recuperar una buena estimación de la imagen de HR potencial mediante la inversión del proceso de degradación que se muestra en la Ec. (1), que puede formularse como

$$\hat{\mathbf{X}} = R\left(\mathbf{Y}, \theta_R\right)$$

donde $R(\cdot)$ representa la función SR y θ_R es el conjunto de parámetros correspondiente. $\hat{X}$ es la imagen superesuelta de $\mathbf{Y}$, es decir, una estimación de la imagen HR real $\mathbf{X}$.

Aparentemente, el proceso de SR y el proceso de degradación son inversos entre sí. Por lo tanto, para obtener un excelente rendimiento de reconstrucción, la función SR $R\left(\mathbf{Y}, \theta_R\right)$ debe adaptarse a la degradación $D\left(\mathrm{X}, \theta_D\right)$. En la literatura, algunos investigadores aproximan la degradación a través del desenfoque,

reducción de resolución e inyección de ruido. Matemáticamente, el proceso de degradación simulado es el siguiente

$$\mathbf{Y} = \mathbf{SBX} + \mathbf{n}$$

donde **B** y **S** denotan las operaciones de desenfoque y reducción de resolución, respectivamente. En general, el desenfoque se realiza convolucionando la imagen HR con un kernel gaussiano. n representa el ruido aditivo, que generalmente se supone que es ruido blanco gaussiano o se puede ocupar un modelo de degradación más simple, es decir, reduciendo directamente la escala de una imagen HR utilizando el núcleo "bicúbico"para generar la imagen LR correspondiente, este ultimo método fue el que se implemento en este proyecto.

5. Descripción del proyecto

5.1. Exploración de los datos

CelebFaces Attributes Dataset (CelebA) es un conjunto de datos de atributos faciales a gran escala con más de 200.000 imágenes de celebridades. Las imágenes de este conjunto de datos cubren grandes variaciones de pose y desorden de fondo. El conjunto de datos cuenta con 202,599 de imágenes con una resolución nativa de 178x128 píxeles.

Además se uso el dataset STL-10 Image Recognition el cual cuenta con 100,000 imágenes con una resolucion de 96x96 y cuanta con 10 clases las cuales son: airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck.

Para la partición de los datos se usaron 4000 imágenes para prueba, 211,820 para evaluación y 86,799 para el entrenamiento.

5.2. Preparación de los datos

Para el preprocesamiento de las imágenes primero se normalizaron cada una de ellas después, se recortaron a 120x120 píxeles cada una, este conjunto son las imágenes HR (X). Para obtener las imágenes LR (Y) se escalaron las imágenes a una resolución de 40x40 píxeles, posteriormente se volvieron a escalar a 120x120 píxeles para así obtener las imágenes con ruido.

5.3. Métodos y modelos

El modelo de autoencoder recibe como entrada una imagen de baja resolución (LR) y su salida es una imagen reconstruida (HR) una vez obtenida esta imagen se calcula la perdida comparándola con la imagen original mediante usando la métrica mean squared error (MSE).

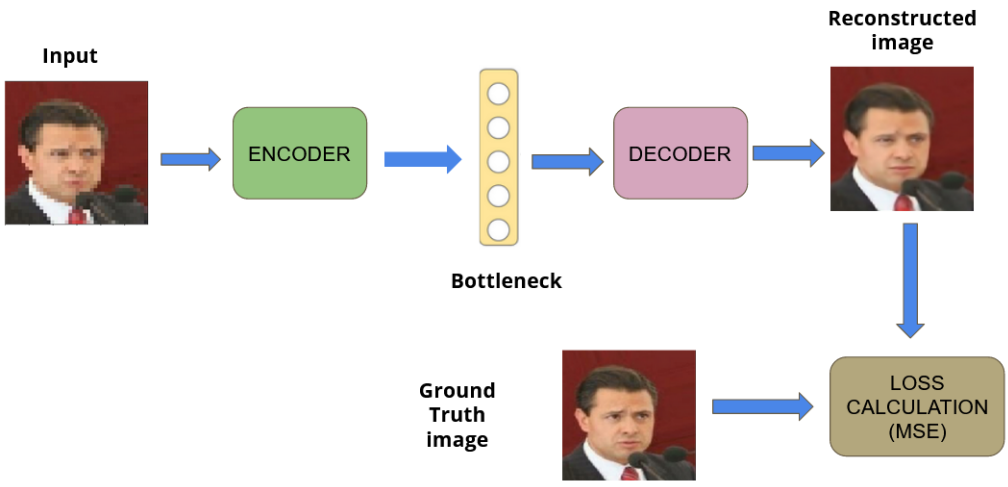

Figura 1. Esquema Autoencoder

Para el **codificador** se ocuparon dos bloque convolucionales, cada uno consta de dos capas convolucionales y un Max pool con filtros de 3x3 y función de activación ReLu, estos bloques se pasan a una ultima capa convoluacional para así formar el codificador completo.

El **decodificador** consta de dos bloques cada uno con una cada de sobre muestreo y dos capas convoluciones, estos bloques pasan a una ultima capa convolucional para así conformar el decodificador.

Por ultimo al modelo se le agregaron dos **conexiones residuales** uno entre la capa 5 y la 9 y otro entre la capa 2 y 13. Las conexiones residuales aportan información extra de la imagen además ayuda contra el desvanecimiento del gradiente. El modelo cuenta con 1,110,403 parámetros.

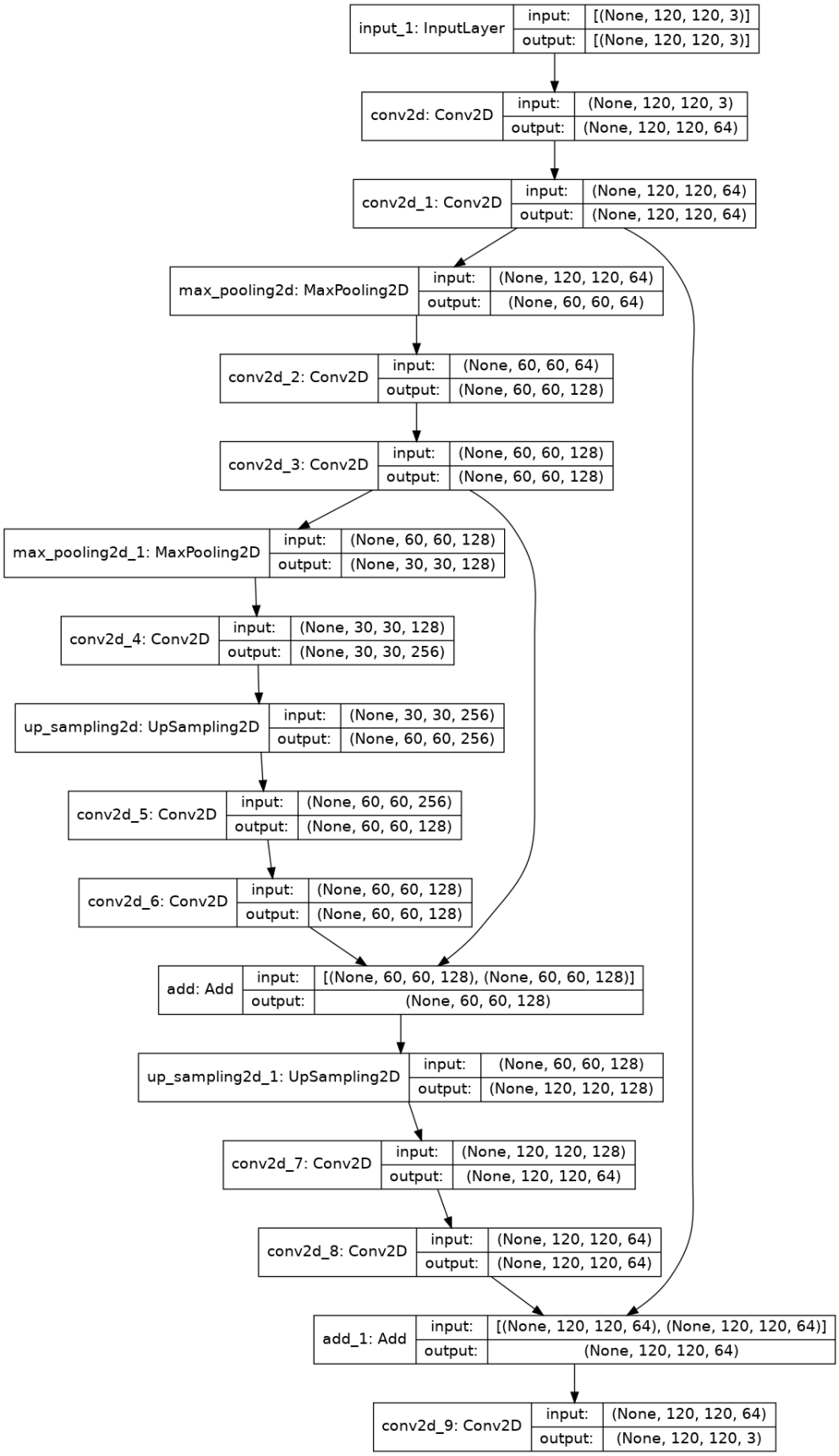

Figura 2. Arquitectura Autoencoder con conexiones residuales

5.4. Evaluación de los modelos

Para aprender el mapeo de imágenes corruptas a imágenes limpias se necesita estimar los pesos Θ representados por los núcleos convolucional y deconvolucional. Específicamente, dada una colección de N pares de muestras de entrenamiento $\left\{X^i, Y^i\right\}$, donde X^i es una imagen ruidosa y Y^i es la versión limpia como la verdad básica. Minimizamos el siguiente error cuadrático medio (MSE):

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \left\|\mathcal{F}\left(X^i;\Theta\right) - Y^i\right\|_F^2$$

El modelo se entreno durante 10 épocas y se ocupo un optimizador Adam y la metrica de accuracy. Los resultados del entrenamiento se muestran en la siguiente figura.

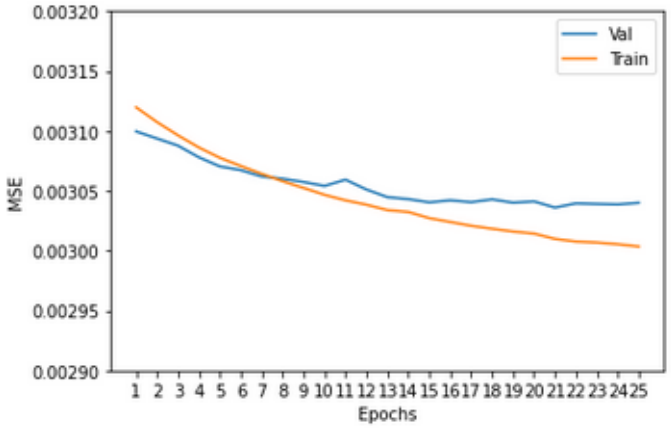

Figura 3. Perdida durante el entrenamiento

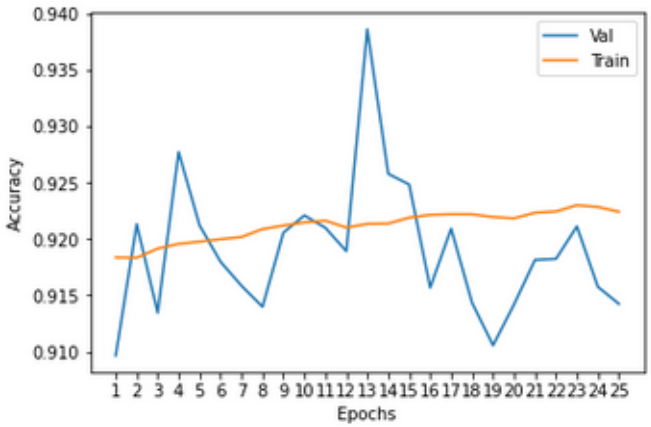

Figura 4. Accuracy durante el entrenamiento

6. Resultados

A continuación se muestran algunos resultados del conjunto de prueba. Podemos observar que a pesar de que la imágenes de baja resolución cuentan poca información y mucho ruido el modelo fue capaz de reconstruir muy bien las imágenes.

Figura 5. Resultados del conjunto de prueba

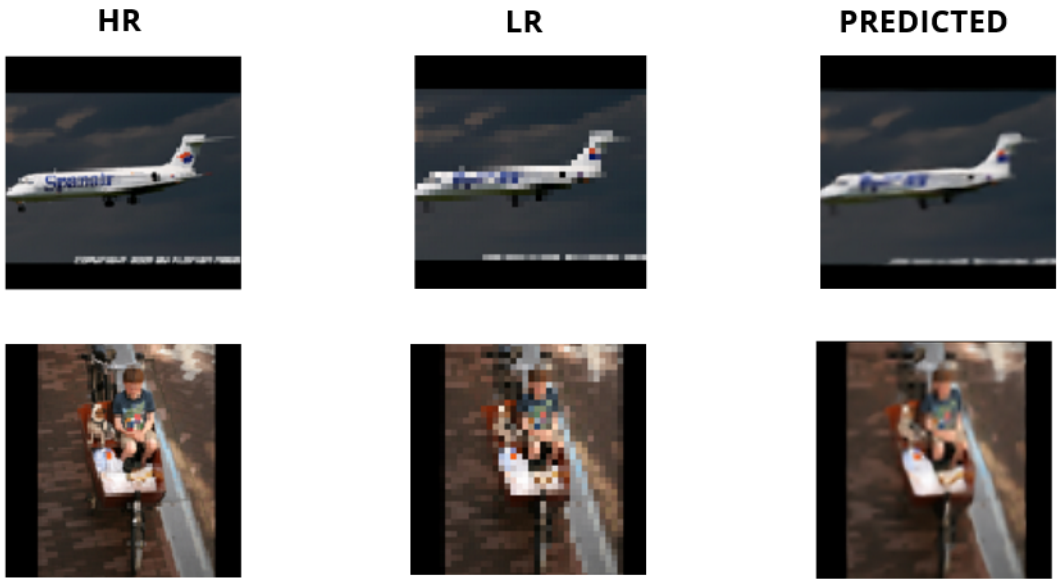

Figura 6. Resultados del conjunto de prueba

7. Conclusiones y recomendaciones

El uso de autoencoders ha mostrado ser un buen método para la reconstrucción de imágenes de rostros. La función de perdida MSE

Referencias

1. Image Resolution Enhancement Using Convolutional Autoencoders
`https://sciforum.net/manuscripts/8259/manuscript.pdf`
2. Unsupervised Real Image Super-Resolution via Generative VariationalAutoEncoder
`https://arxiv.org/pdf/2004.12811.pdf`