

LSTM AND CONVOLUTIONAL NETWORKS FOR RECREATING MEXICAN “ANTOJITOS” IMAGES



Victor Hugo Martinez Huicochea¹

Escuela Superior de Física y Matemáticas

Lourdes Fabiola Uribe Richaud¹

Instituto Politécnico Nacional

Abstract

Currently, technology is rapidly developing. As a result, many technologies had arisen in many fields, standing out computational techniques such as neural networks. Inspired by the human brain processes, neural networks allow us to answer real life questions.

In this work, neural networks were implemented to process text describing Mexican “antojitos” and generate images accordingly to the text to assess their performance. This work aims to provide the foundation for the development of complex tasks whether it is reproducing and correcting blurred images or improving communications through visual aids.

Theoretical Background

Neural networks are an automatic learning model that processes information by simulating biological neurons, consist in three main parts: input layers, where information is received, hidden layers, which process information by applying multiple regressions through their respectively activation functions, and the output layers that provide the result of the model.

Convolutional neural networks are used for processing pictures, searching for crucial elements by applying characteristic detection also known as kernel or filter. Some other methods such as matrix convolution and gathering are used.

The Long-Short Term Memory (LSTM) network is a Recurrent Neural Network (RNN) type, where the output is used as an extra input to feedback future instances. LSTM networks can have many gates mainly the next ones: the forgetting gate, which removes unnecessary information, input gate that decides which values to keep and output gate, which updates the hidden value and show the given result.

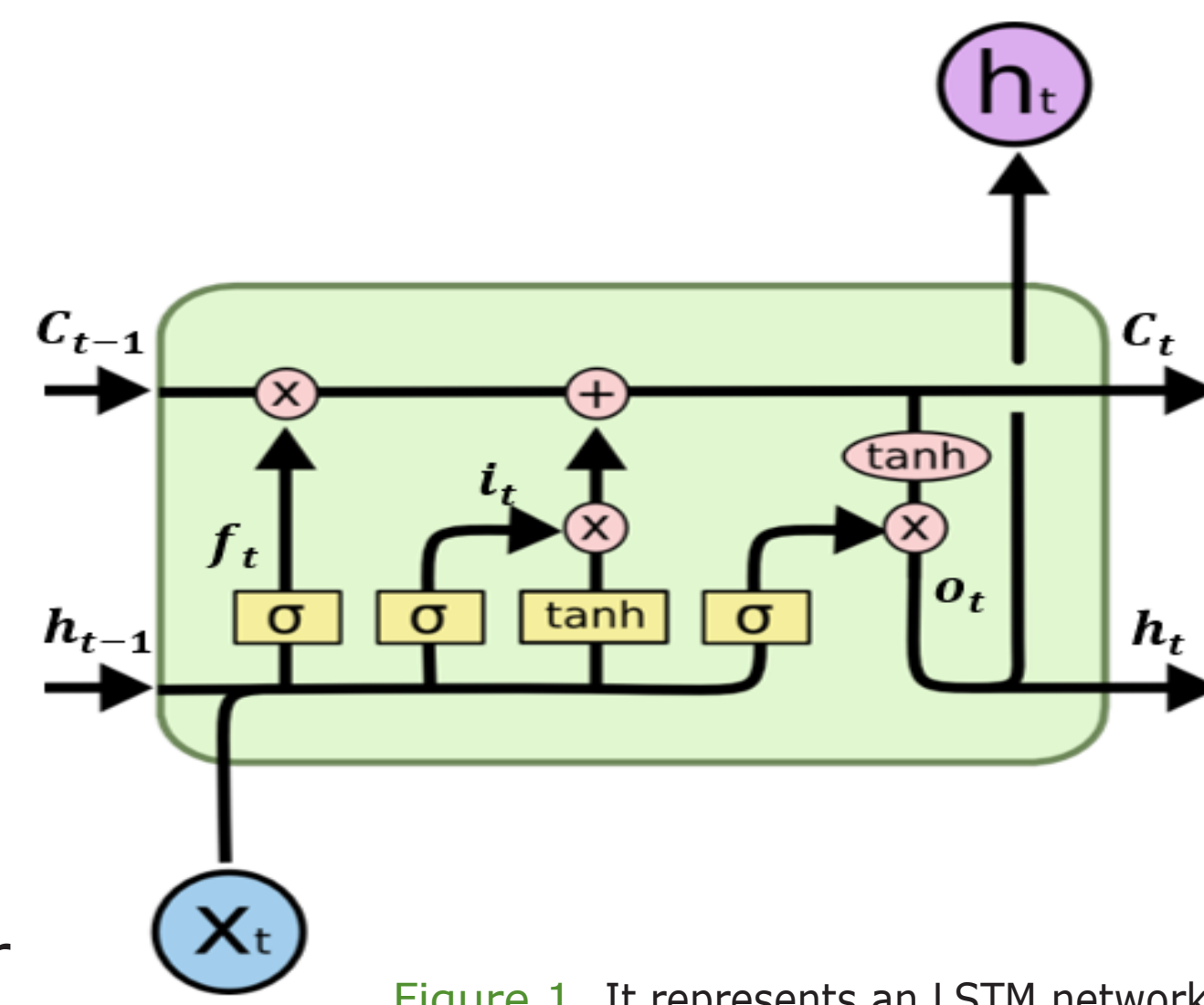


Figure 1. It represents an LSTM network where the gate (C) is affected by the hidden value (h) and the forgetting gate (f), the input (I) and the output (O).

Creation of the Data Base

For training the model, a database was created using version 1.1.2 of the “bing-image-downloader” library to export 100 images of different “antojitos”, which is listed on table 1, to supervise the learning process.

Antojitos
Tacos Campechanos
Enchiladas
Mole Poblano
Gorditas de Nata
Quesadillas
Tamales
Chiles en Nogada
Sopes
Pozole
Burritos

Table 1. The selected “antojitos” used in the learning process.

Also, the database was cleaned so finding blurred images for the deep network training. After that, a brief description about the dish was provided to the network input. The finale text was formed by 784 different words in total, distributed with a maximum of 25 words per description.

Then, using TensorFlow library and the ImageDataGenerator function, a set of the next functions was randomly applied: rotation, cut, increment and mirrored; alongside the original labels incremented from 1,000 to 10,000 instances.

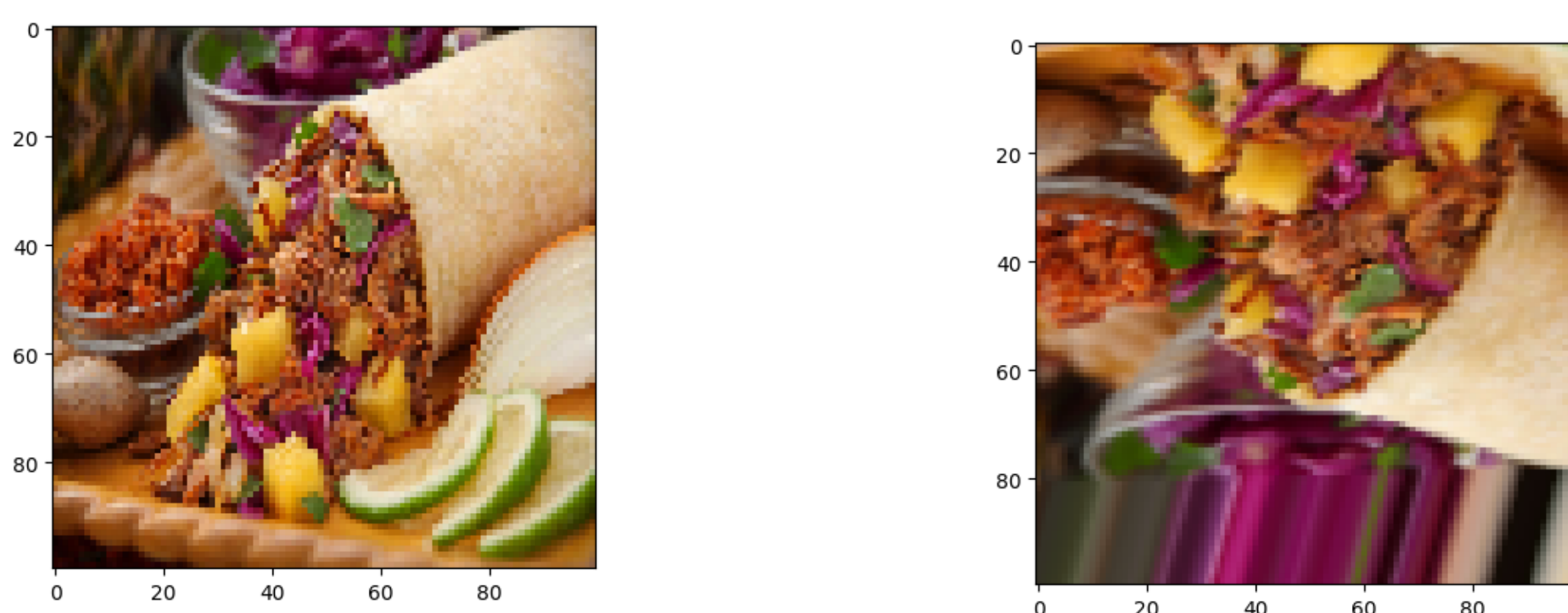


Figure 2. A burrito image, the original image on the left and on the right after the function ImageDataGenerator was applied. Refer to: Hacienda Corona (2023). Burrito. Extracted: June 12, 2024, from: <https://www.haciendacoronama.com>

Training

Firstly, all the food labels were exported and stablished, and with the function Tokenzier from keras a question bank was created, in which the phrase were transformed into vectors, Then, the matrices were adjusted to a 100x100x3 sized for a better management, the dimensions of the matrices were the references of length, width and the RGB color code of the image. Consequently, the values were normalized to improve the data processing. A simple architecture was used to optimize the processing time. Table 2 shows the architecture.

Layer (type)	Output Shape	Param #
embedding_3 (Embedding)	(None, 25, 1200)	940800
lstm_6 (LSTM)	(None, 25, 1200)	11524800
lstm_7 (LSTM)	(None, 25, 1200)	11524800
reshape_3 (Reshape)	(None, 100, 100, 3)	0
dropout_3 (Dropout)	(None, 100, 100, 3)	0
conv2d_6 (Conv2D)	(None, 100, 100, 3)	3603
max_pooling2d_3 (MaxPoolin g2D)	(None, 100, 100, 3)	0
conv2d_7 (Conv2D)	(None, 100, 100, 3)	228
Total params: 23994231 (91.53 MB)		
Trainable params: 23994231 (91.53 MB)		
Non-trainable params: 0 (0.00 Byte)		

Table 2. Network Model is used for image generation.

A two layers LSTM was implemented allowing the network to learn about grammar and the relation between words and food, the all the elements of the 100x100x3 dimension matrices were reorganized to be used as the starting point for our RGB image and pass through two layers convolutional networks.

The network was trained with 10,000 stages by the Nadam optimizer and the Mean Squared Error loss function, using 80% of the data and the other 20% for validation. It needs to be emphasized that every 500 stages backups were made.

Result

After running the network, some images were created obtaining the next results:

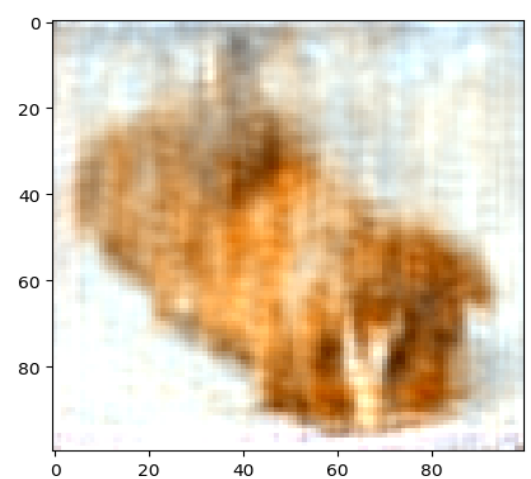
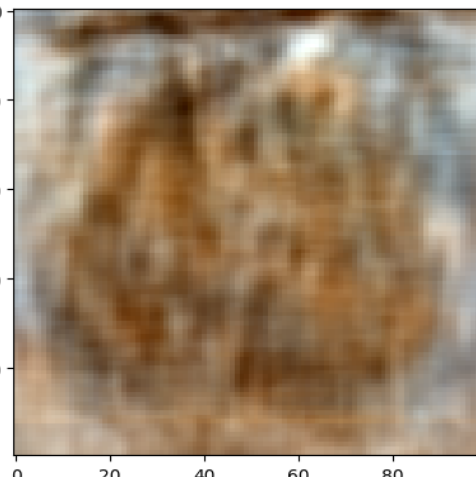
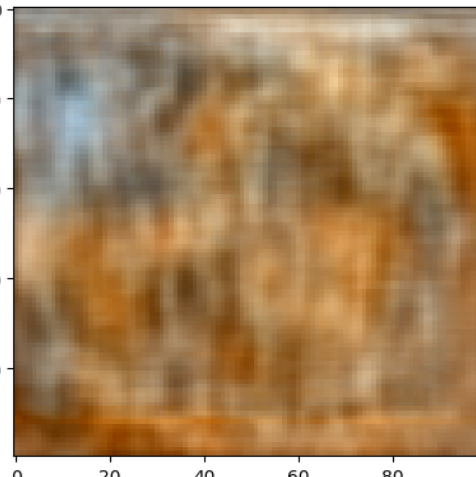
phrases given	generated images
“taquitos con papas y cebolla rica”	
“mole poblano con arroz”	
“gorditas de nata rellenas”	

Table 3 Some of the neural network results.

As it is shown in Table 3, the generated images are blurred and lack precession. However, it can also generate similar images to what was requested. For example, it generates a rounded yellowish image when tacos were requested, in case of mole, the color is brown with dark red shades, also, notice that the network has some difficulties generating the “gordita de nata.”

In this project, we can appreciate the performance of LSTM and convolutional neural networks in the creation of images under limiting conditions, such as a reduced database, hardware and software limitations and a basic architecture, obtaining a tool that can identify y create simple forms.

How to improve the network?

Many aspects of the neural network can be improved, some of the questions asked about its development would be: How would the performance be affected if it had a larger architecture? By increasing the image size, would the training and output improve? What happens if it is trained with a richer and bast database? What happens if a diffuse model was to be added? What if neural networks based on stochastic processes were added?

Those questions would help develop the network and complement it.

References:

- Hacienda Corona (2023). Burrito. Extracted: June 12, 2024. Site: <https://www.haciendacoronama.com>
- IBM. (2024, may). ¿Qué es una red neuronal? Extracted: August 23, 2024. Site: <https://www.ibm.com/mx-es/topics/neural-networks>
- IBM. (2021, december). ¿Qué son las redes neuronales convolucionales? Extracted August 26, 2024. SItE: <https://www.ibm.com/mx-es/topics/convolutional-neural-networks>
- Olah, C. (2015, august 27). The repeating module in an LSTM contains four interacting layers. Colah’s Blog. Extracted: August 23, 2024. Site: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- Singh, G. P. (2022). Bing-Image-Downloader (1.1.2) [Software]. <https://pypi.org/project/bing-image-downloader/>