Video Super Resolution Using Neural Networks

Dominik Chodounský

29th November 2021

1 Overview

The aim of this project is to take a short low-resolution video and use deep learning techniques to upscale the video resolution via synthetic generation.
More specifically, the project should explore the usage of the U-Net and GAN architectures to achieve this goal.
I will begin by looking into techniques for single image super resolution and then try to transfer those models to the video domain.

2 Research

This section summarizes the articles I have looked into as part of my ongoing research on this topic.

- Image Super-Resolution Using Deep Convolutional Networks

  In this article, the authors create an end-to-end mapping between low-resolution and high-resolution images via a feedforward deep convolutional neural network (CNN). The authors break the upscaling process into three parts. The first is patch extraction where patches are extracted from the low-res image and then represented as a higher-dimensional vector. The next part performs a non-linear transformation of this vector and the final part aggregates the high-dimensional patches to construct the final high-res image. The evaluation of the result is by comparing it to how similar it is to ground truth (original high-res image before it was downscaled to serve as input). The authors used mean squared error (MSE) as a loss function, which favours a high peak signal-to-noise ratio (PSNR) which is commonly used to evaluate image restoration tasks and is partially related to the perceptual quality, which is often hard to grasp mathematically. Another evaluation metric used was structural similarity index measure (SSIM).

- U-Net: Convolutional Networks for Biomedical Image Segmentation

  This article presents the U-Net model, which is a CNN architecture consisting of a contracting and an expanding path, which are interconnected by skipping connections. The contracting path is made up of convolutional and maxpooling layers and the expanding path performs upsampling with deconvolutions. The model is commonly used for image segmentation tasks in the medical domain, but there are several use cases of it being used for super resolution tasks due to its relatively low training data demands and precise feature localization.

- RUNet: A Robust UNet Architecture for Image Super-Resolution

  The authors of this article present an improved architecture for performing single image super resolution with an improved U-Net architecture, which they call the RUnet (Robust U-Net). The individual steps in the contraction path are expanded into residual blocks so that the network is able to learn more complex structures. They also use a perceptual loss function, where instead of comparing the output and the ground truth pixel-wise, they feed them both into a convolutional feature extractor (commonly the VGG networks) and measure the distance between the two mapped images in the feature space.

- Deep Recurrent Resnet for Video Super-Resolution

  In order to expand the task from single image super resolution to the video domain, there are various approaches. One such approach uses optical flows to handle video frames, which attempt to compensate for the motion and perform multi-frame super resolution. The authors of this paper describe how this approach creates an inefficient bottleneck and suggest and improvement that takes advantage of the recent advances in recurrent neural networks, more specifically the long-short-term memory (LSTM) network. By using convolutional LSTM modules, they achieved superior reconstruction of videos that has temporal coherency.

The following articles have yet to be researched properly...

- *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*
- *Image and Video Super Resolution using Recurrent Generative Adversarial Network*

3 Technologies

The implement of my experiments uses the high-level Keras functional API based on the Tensorflow framework. So far, all experiments were done in a Jupyter Notebook environment on my local machine (*Intel Core i7-9750H* processor, 16 GB DDR4 RAM and an *NVIDIA GeForce GTX 1650* graphics card with 4 GB of VRAM), but I plan on using Google Colaboratory for larger models that have larger memory requirements.

4 Data

I plan on using the Vimeo-90K Dataset for training and testing my models. It is a large and diverse dataset of video clips from the *Vimeo* hosting platform. More specifically, I will use the provided septuplet version of the dataset which includes 91 701 7-frame video sequences extracted from 39 000 different video clips from the Vimeo-90k. This dataset was specifically designed for video denoising, deblocking and super resolution tasks.
Due to the large size of the dataset (82 GB), I am waiting for better internet qualityin order to be able to download it in a reasonable time frame. Therefore, in the meantime, I tried experimenting with single image super resolution on the famous

Dogs vs. Cats dataset, which has a more managable size. The dataset is already split into a 25 000 training images and 12 500 test images.

5 Progress

As of writing this milestone report, I have experimented with building my own adapted version of the U-Net network and performing single image super resolution on the Dogs vs. Cats data. The architecture takes 64 $\times$ 64 pixel RGB images, then repeatedly downscales them in the contracting path with convolutional and maxpooling layers, and then performs the upscaling in the expanding path via deconvolutions and filter concatenations with corresponding layers in the contracting path. After reaching the output size of 64 $\times$ 64 pixels, there is a final layer added on top of the U-Net which performs the final upscaling to produce a 128 $\times$ 128 image trained to have super resolution. I have called my adaptation of the architecture the SR-U-Net.
The described SR-U-Net was then trained for 10 epochs with the Adam optimizer (initial learning rate of 0.001), using mean squared error loss as the optimization criterium. During the training, I also observed the PSNR and SSIM metrics, which I will later use for comparison with other training attempts.
Aside from the training image data generator used during the trianing loop, I also implemented a test generator to evaluate the model predictions with. The results can be seen in Figures 1-6 at the end of this report, where we see the original 128 $\times$ 128 image, then there is a low resolution 64 $\times$ 64 version of the image and also the super resolution version which was created by feeding the low resolution image through the SR-U-Net. As we can see, the resolution upscaling certainly works, as the generated images are much smoother, more easily recognizable and natural to the human eye. We do, however, see a significant different in image sharpness in comparison to the original high-res images. This is true for all test cases, so I will try to experiment with other models which could perhaps provide better results.

6 Planned improvements

Following this milestone report, there are several improvements I have planned for this project. Firstly, I will try to download and start working with the above mentioned Vimeo-90k dataset and start putting the upscaled frames together to form a short video, as that is the goal of this task.
I will try to implement a GAN architecture in order to compare it to the U-Net approach.
In order to try to minimize temporal incoherency, I will try to incorporate recurrent layers into my models, perhaps with the use of LSTM modules as suggested by several articles found during my research.
Finally, I will try various loss functions, such as the perceptive loss achieved by running the compared images through a convolutional feature extractor such as VGG16.

Figure 1: Example of single image super resolution on a cat image.

Figure 2: Example of single image super resolution on a cat image.

Figure 3: Example of single image super resolution on a cat image.

Figure 4: Example of single image super resolution on a dog image.

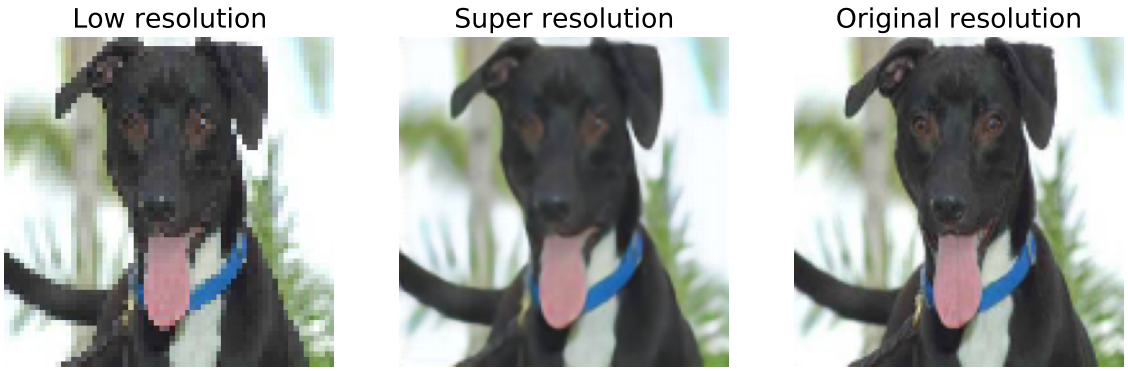

Figure 5: Example of single image super resolution on a dog image.

Figure 6: Example of single image super resolution on a dog image.