Jesus Chavez

ITAI 2376

MD Report

Forward diffusion is, in essence, a method where an image is spoiled by gradually adding a tiny bit of random noise in several steps. In the same time the image fades, it looks a bit noisy. You can follow a perfectly clear photo turns into a noisy one. At each step, the image becomes a bit more blurry and the quality goes down. Through this carefully noise-adding procedure, there is a route leading back to the original image. Rather than fully removing the noise at once, the model gradually sees little changes.

Usually, by cutting the noise by about 30 to 50%, the figures in the image become clearer. While the digits 0 and 3 were very different at around 60, the digit 7 was not very prominent in the sample images. It was due to seven of the hundred steps. The degree of shapes and details of the numbers that each person can see is different.

Due to its encoder-decoder architecture, U-Net is able to see both the overall and the very subtle aspects, hence by this very architecture it is able to effectively handle diffusion while at the same time compressing and restoring features. The architecture layout in question is instrumental in noise reduction.

Skip connections highlight the most important details thus we do not lose the borders of the numbers. Class conditioning with c_mask enabled classifier-free guidance during training simply by converting the digit we want into a spatial feature map using a one-hot vector.

MSE being smaller means that the model is better at removing the noise and predicting it; loss values indicate its performance. We noticed differences during our training; the photos' clarity changed from being vague to bright numbers with the help of time embedding for the purposes of improving the taking of pictures techniques.

CLIP scores indicate the degree of agreement between the images and the text; higher scores mean a better link. Easier numbers, like 1 and 0, get higher scores because they are simpler, whereas more complicated numbers like 5 and 8 do not get that good scores. One can improve the results by selecting the best ones in terms of CLIP scores and considering changing the model with the CLIP-related loss for better accuracy.

Among the real-world applications are the data creation for handwritten numbers, the generation of font styles for digits, and the development of learning tools. The problems come in the form of low diversity, boring style adaptation, and the time taken for the creation. Possible future improvements include guidance without classifier networks, mixed training of different datasets, and DDIM or DPM methods.