Human Perception-based Image Enhancement Using a Deep Generative Model

Amir Nazemi, Shiraz University, Shiraz, Iran
Shima Kamyab, Shiraz University, Shiraz, Iran
Zohreh Azimifar, Shiraz University, Shiraz, Iran
Paul Fieguth, University of Waterloo, Canada

Abstract

In this paper we propose a deep model for perceptual image enhancement based on generative modeling. The proposed framework is inspired by the Conditional Variational AutoEncoder (CVAE) which is a well-known deep generative structure. In generative models, there are efficient regularizers for controlling the output distributions using information from input data which lead to accurate and visually plausible results with few parameters. Additionally, we propose to use an image quality assessment network to determine the best result among those obtained by the implemented CVAEs. The proposed CVAE structure models the histogram vectors of different color channels and parameters of image data (i.e., the networks do not work directly on pixel values). This configuration makes the proposed framework capable of using images of different sizes. Qualitative and numerical evaluations on a related dataset compared to state-of-the-art indicate superiority of the proposed framework in improving image quality and content.

1 Introduction

Image enhancement is a long-standing field in computer vision and refers to a set of techniques (such as filters, editors, etc), which manipulate image pixel values to improve the visual quality of that image. These techniques have been widely applied in computer vision [3].

Before defining an enhancement technique, a metric is required to measure the visual quality of an image. In this paper, the proposed metric is based on human visual perception, that is the metric undertakes an *aesthetic assessment*. Using this metric, the objective of enhancement approaches is to improve aesthetic quality of the image.

Aesthetic assessment approaches are categorized in different ways. From the image features aspect, there are quality assessment methods, which focus on low-level features like noise, compression, artifacts [6], and from the image semantic viewpoint, aesthetic assessment methods deal with emotion and beauty [2]. From the objective aspect, there are reference-based assessment methods [11], which use the target image to measure the visual rank of an image and no-reference methods [8], which evaluate the statistical information of image distortion to measure the image quality.

Image enhancement is an ill-posed and non-linear problem, and we address this problem by using a regularized optimization framework in a constrained solution space. We propose to formulate image enhancement as a generative process. The generative models utilize appropriate strategies on controlling output distribution and embedded regularizers applicable for these structures.

Recently, deep learning-based approaches, as powerful representation methods, have widely addressed many problems including image enhancement [11, 4]. However, existing solutions often are problem specific and there is still a need for quality and interpretability improvement in this field. In this paper, a deep generative structure called CVAE [5]is used for the image enhancement optimization problem. We believe the regularizer and control strategy on the distribution of latent variable embedded in this structure, make it a powerful optimization baseline for image enhancement.

Therefore, the main contribution in this paper is two-fold. We design a specific deep generative model (CVAE) to address image enhancement problem. Also, we propose to reduce the network parameters by using new features for image enhancement including the histogram of RGB and LAB images and four selected pixel values as input to the network, instead of raw images.

This paper is organized as follows: Sec. 2 reviews some of the most recent related works in the field of quality assessment and perceptual image enhancement. In Sec. 3, we formulate the image enhancement problem as a generative model and propose our framework based on the CVAE structure. Some comparative numerical and qualitative results are conducted in Sect. 4 to evaluate the performance of proposed framework compared to state-of-the-art. Finally the paper is concluded in Sec.5.

2 Related Works

As a state-of-the-art method in the field of image quality enhancement, we refer to [11] with a structure consists of two CNN modules. The fixed weights of the first trained CNN are used as a regularizer function and is considered as a part of a new loss function. The second CNN is trained with the proposed loss function on the MIT-Adobe FiveK dataset [1]. In [4], image enhancement parameters are learned through adversarial learning driven by aesthetic judgment. This framework permits multiple forms of image enhancement and is capable of encapsulating various functions such as scaling, translation, contrast adjustment, and color enhancement in a unified network. In [7] a so called *retouching* framework is proposed as a *white box* using reinforcement learning and adversarial training, with the aim of providing a sequence of actions on the image with the preference of user.

Conditional Variational Autoencoder (CVAE), based on the Variational Autoencoder (VAE) first proposed in [9], is a deep generative latent variable model for implicitly estimating the underlying distribution of the data. In its basic form, it has a stochastic architecture directly controlling the latent variable distribution using the information from the data and some form of regularizer, i.e., *KL-divergence*, in its objective function.

The existing studies about image enhancement using deep learning could achieve plausible results in solving image enhancement problem. We believe proposing a generative model based frameworks like GANs [4] and CVAEs [9], for solving such problems leads the solution to be well formulated, since such formulation captures more properties of this problem by adopting appropriate regularization and controlling output distribution and therefore improves the results in presence of relatively low parameter setting.

In the next section we demonstrate the formulation and the structure of the proposed network in this paper.

3 Proposed Framework

In this section, we first formulate our proposed CVAEs based on the basic formulations of VAE [5], then explain our LAB color enhancement module, finally our proposed framework structure.

3.1 CVAE Formulation

As demonstrated in [5], the Variational AutoEncoder (VAE) is a latent variable generative model with the objective of maximizing the likelihood of training data, assuming a latent variable to produce them:

$$\max_{\theta} P(X) = \int P(X|z;\theta)P(z)dz \tag{1}$$

where θ denotes the system parameters to be identified, z is the latent variable and X is the training data fed to the system. In the case of the image enhancement problem we denote X as the target reference image.

In this paper we propose to use $Q(z|Y;\beta)$ as encoder's output distribution, where the information to control latent variable distribution, Y, can be from sources other than X. More formally, the output of VAE is $E_{z\sim Q}(P(X|z;\theta))$, where the latent variable distribution, i.e., $Q(z|Y;\beta)$, is constrained to be like the true probability $P(z|X)$, i.e., the probability of inferring z from X. This constraint is formulated using KL-divergence operator:

$$D(Q(z|Y;\beta)||P(z|X)) = E_{z\sim Q}[logQ(z|Y;\beta) - logP(z|X)] \tag{2}$$

Applying Bayes rule to $P(z|X)$ and changing term, we have:

$$\begin{aligned} logP(X) - D(Q(z|Y;\beta)||P(z|X)) = \\ E_{z\sim Q}[logP(X|z;\theta)] - D(Q(z|Y;\beta)||P(z)) \end{aligned} \tag{3}$$

where β and θ are parameters of trainable encoder and decoder, respectively.

The left-hand side in (3) is equivalent to the objective of VAE to be maximized and the right hand-side can be represented by VAE. In the VAE objective it can be observed that the KL-divergence term appears like a regularization term, which forces the distribution of the latent variable to be obtained from the data itself. It is a generic regularization term that can be used in any optimization problem and we show in Sec. 4 that it is suitable for the image enhancement problem. Therefore, VAE can be considered as a regularized optimization, which is desirable in complicated optimization problems. It is worth noting that $P(z)$ is set to $N(0,I)$ so that it can be converted to any distribution consistent with the latent variable in the trainable decoder.

In the case of CVAE, all distributions are conditioned to some measurement, called C, and this condition is used as input to both encoder and decoder components.

In this paper, as above formulation indicates, we design a framework consisting of multiple CVAE-based components, having above formulation, in which the input to encoder Y and output of decoder X are not the same. The former are RGB and LAB histograms of image to be enhanced and the latter is the RGB and LAB histograms of target image. We set the condition to be the input to encoder Y plus some information about the image in the form of binary label defining the image capture properties like light condition, location, etc. Using histograms instead of image pixels, the system will depends only on few parameters and does not have to deal with different sizes of input images.

In the proposed CVAE-based structure, the encoder component is not omitted in the test phase in order to obtain appropriate samples from $Q(z|Y;\beta)$. Having defined the proposed framework formulation, next, we will describe our proposed LAB color module, which provides parameters for two of the CVAE-based components in our framework. There is, also another CVAE-based component, which directly works with RGB channel of input and target images.

3.2 LAB Color Enhancement

For enhancing the color contrast of a picture in LAB color, which is designed to model human visual perception, the experts follow some rules to alter the L, a and b channels of an input image. The adjustment function, which changes the pixels' intensity of a and b channels, is described in [10] as:

$$f(m) = \begin{cases} 0 & \text{for} \quad m < \alpha \\ \frac{255}{\beta-\alpha}(m-\alpha) & \text{for} \quad \alpha \le m \le \beta \\ 255 & \text{for} \quad m > \beta \end{cases} \tag{4}$$

where α and β are estimated using two pixel values from an input channel and its mapping into the output image channel. We formulate this estimation as follows:

$$T = \frac{\psi(m_2)-\psi(m_2)}{m_2-m_1} \tag{5}$$

$$\alpha = m_1 - \frac{1}{T}\psi(m_1) \tag{6}$$

$$\beta = \frac{255}{T} + \alpha \tag{7}$$

In the above equations, m_1 and m_2 are two pixel values of the input image and $\psi(m_1)$ and $\psi(m_2)$ are the mapped pixel values of m_1 and m_2 in the enhanced target image. In our dataset, this mapping is done by an expert and we assume this operation in the form of a piecewise linear function ψ. In other words, we estimate the parameters of function f assuming that functions ψ and f are equal. T is the slope of function ψ and by making the slopes of two Piecewise linear functions f and ψ equal, we can derive (6) and (7). Figure 1 shows the plot of function $f(m)$ and its parameters on histogram of channel a of an input and histogram of a channel of output image.

3.3 Framework Architecture

In the proposed network structure there are three CVAEs, with specific representations of image to be enhanced as input and its human expert enhanced image pair as output.

In all our CVAE models we convert the the semantic information about each photo into a one-shot vector and use it as a condition

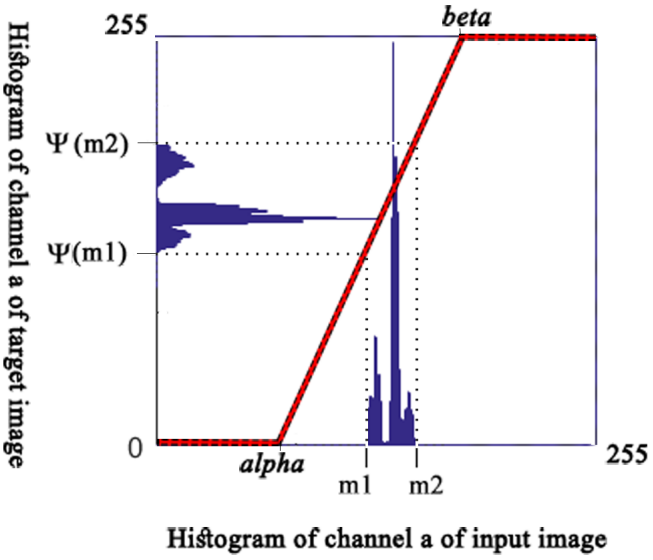

Fig. 1: The plot of Eq. 4 (solid red line). The parameters of $f(m)$ on the channel a of an input image in the LAB color space, can be estimated by four pixel values including m_1 and m_2 of input image histogram and $\psi(m_1)$ and $\psi(m_2)$ of output image histogram. It is assumed that $f(m)$ and $\psi(m)$ are equal.

for our CAVE models. It helps to have different latent variable distributions for each semantic label. Additionally, since our dataset is small and highly variable, the input vector is short-circuited to the input of the decoder to have a better transferring histogram.

In our proposed framework we use three different kinds of features from the input. We choose the RGB and Lab color spaces of input images for enhancement. The RGB channels are robust to big changes. Furthermore, The LAB color space is the favourite color space among photographers for editing photos.

We extract three image histogram vectors of the RGB channels of the input image and concatenate them to the condition vector and feed them to the encoder. We consider the concatenation of the RGB histogram vectors of our target enhanced image pair as the output of this network. The second CVAE works on the first channel of the LAB channels of the input image and learns how to transfer the input histogram of L channel into the enhanced output's histogram of its L pair channel.

As explained in the previous section, for enhancing the a and b channels of the input LAB color we need to calculate the parameters α and β for channels a and b. Here, since our dataset does not provide the parameters α and β, we measure these two parameters using two pixel values of the input image and their mapped values into the target image pair. We use the starting and ending point of each channel's histogram of the input and target images as $m_1, m_2, \psi(m_1)$ and $\psi(m_2)$ and use (5), (6) and (7) to calculate α and β. We employ (4) to enhance the channels a and b. In other words, a photography technique is modeled for image enhancement.

For CVAEs whose output is a histogram (RGB or L), we use histogram matching to adjust the input image, in which the histogram of the input channel be similar to that of the target image. This strategy has several benefits for our structure such as being independent from the input image size. In the last step, we employ the NIMA network [12] to choose the most perceptually beautiful image between the outputs of the CVAE model on the RGB channels and photography inspired models on the Lab channels.

4 Experimental Results

The MIT-Adobe FiveK dataset [1] contains 5000 raw pictures captured by SLR cameras and five enhanced images for each input prepared by five experts. For this article we use 4800 images as training data and 200 images for validation and test. We choose one expert (expert C) among five expert's output as the target for our experiments.

Figure 3 illustrates the box-plot of NIMA scores of the input, target and our model's enhanced output data. Each of our CVAE models on RGB and and on the LAB color improves the NIMA scores of the input data and enhances the input. Furthermore, as is seen in