Colorful Image Colorization

Richard Zhang, Phillip Isola, Alexei A. Efros

https://arxiv.org/abs/1603.08511
https://arxiv.org/pdf/1603.08511.pdf

TL;DR

Paper proposes a system to imagine a plausible colour version of a grayscale photograph. Previous approaches relied on user interaction or resulted in bad colourizations. This fully automatic approach produces vibrant and realistic colourizations. The colourization problem is quantified as a classification task and uses a CNN. This method fooled humans 32% of the time in a colourization Turing test.

Introduction

Generating colour may seem daunting, but often, the semantics of the scene and texture provide enough information to predict colour. The task is therefore to model the statistical dependencies between the semantics and the textures of grayscale images and their colour versions.

Any colour photo can be used for training. The image's *L* channel is used as the input, and its *ab* channels as the supervisory signal. Previous attempts at the problem often produce desaturated results. A tailored loss function is used to counteract this. As colour prediction is multimodal, we predict a distribution of possible colours for each pixel, and emphasize rare colours to exploit the diversity of large-scale data(1M+ photos). The final colourization is produced by taking the annealed-mean of the distribution.

Evaluation is done through a colourization Turing tests, where participants identify the fake colourization.

Approach

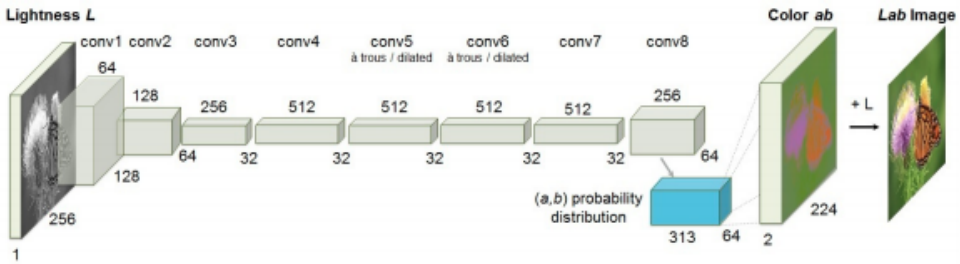

Fig. 2. Our network architecture. Each `conv` layer refers to a block of 2 or 3 repeated `conv` and `ReLU` layers, followed by a `BatchNorm` [30] layer. The net has no `pool` layers. All changes in resolution are achieved through spatial downsampling or upsampling between `conv` blocks.

A CNN is trained to map from a grayscale input to a distribution over quantized colour value outputs. Given an input lightness channel, the objective is to learn a mapping $\hat{\mathbf{Y}}=\mathbf{F(x)}$ to the two associated colour channels. The CIE *Lab* colour space is used. Euclidean Loss can be used to measure the distance between predicted and ground truth colours. However, the optimal Euclidean loss would be the mean of the set, and in colour prediction, this averaging effect favours gray colours. Instead, the problem is treated as one of classification. The *ab* output space is quantized, and for a given **X**, we learn the mapping $\hat{\mathbf{Z}} = \mathcal{G}(\mathbf{X})$ to a probability distribution over possible colours. To compare the predicted colour against ground truth, cross entropy loss is used.

The distribution of *ab* values in natural images is biased towards colours of clouds, pavement, dirt, and walls. To account for this, the loss is reweighted based on the pixel colour rarity.

$$\mathrm{L}_{cl}(\hat{\mathbf{Z}}, \mathbf{Z}) = -\sum_{h,w} v(\mathbf{Z}_{h,w}) \sum_{q} \mathbf{Z}_{h,w,q} \log(\hat{\mathbf{Z}}_{h,w,q})$$

v() is the weighting term to rebalance loss

$$v(\mathbf{Z}_{h,w}) = \mathbf{w}_{q^*}, \text{ where } q^* = \arg\max_q \mathbf{Z}_{h,w,q}$$

$$\mathbf{w} \propto \left((1-\lambda)\tilde{\mathbf{p}} + \frac{\lambda}{Q}\right)^{-1}, \quad \mathrm{E}[\mathbf{w}] = \sum \tilde{\mathbf{p}}_q \mathbf{w}_q = 1$$

Finally, we define *H*, which maps the predicted distribution to point estimate in *ab* space. To do this, the annealed-mean is taken of the distribution.

$$\mathcal{H}(\mathbf{Z}_{h,w}) = \mathrm{E}\big[f_T(\mathbf{Z}_{h,w})\big], \quad f_T(\mathbf{z}) = \frac{\exp(\log(\mathbf{z})/T)}{\sum_q \exp(\log(\mathbf{z}_q)/T)}$$

T=0.38 is used in the paper. *H* operates on each pixel independently

Experiments

The network was trained on the 1.3M images from ImageNet.