The Contextual Loss for Image Transformation with Non-Aligned Data

Mengjie Sun
2018.7.25

Introduction

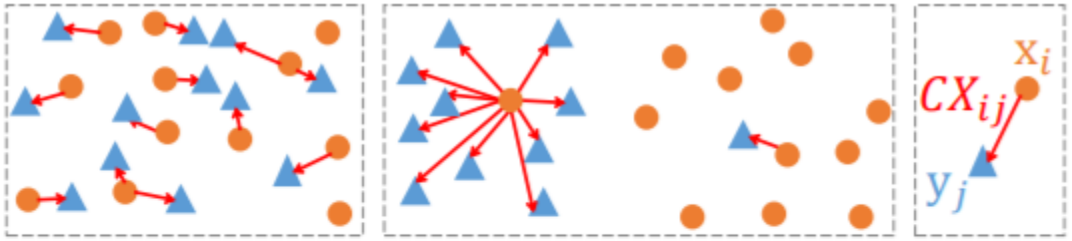
Feed-forward CNNs trained for image transformation problems rely on loss functions that measure the similarity between the generated image and a target image. Most of the common loss functions assume that these images are spatially aligned and compare pixels at corresponding locations. However, for many tasks, the aligned training pairs of images will not be available. This paper presents an alternative loss function that does not require alignment, thus providing an effective and simple solution for a new space of problems.

Non-aligned training data

Related Work

Image-to-Image Translation
Domain Transfer
Style Transfer
Semantic Style Transfer

Method



(a) Similar (b) Not-similar

Fig. 3. Contextual Similarity between images: Orange circles represent the features of an image x while the blue triangles represent the features of a target image y. The red arrows match each feature in y with its most *contextually similar* (Eq.(4)) feature in x. (a) Images x and y are similar: many features in x are matched with similar features in y. (b) Images x and y are not-similar: many features in x are not matched with any feature in y. The Contextual loss can be thought of as a weighted sum over the red arrows. It considers only the features and not their spatial location in the image.

The Contextual loss

$$\mathcal{L}_{\mathrm{CX}}(x, y, l) = -\log\left(\mathrm{CX}\left(\Phi^l(x), \Phi^l(y)\right)\right)$$

Define the contextual similarity CXij between features xi and yi:

$$\tilde{d}_{ij} = \frac{d_{ij}}{\min_k d_{ik} + \epsilon} \qquad \text{fixed } \epsilon = 1\mathrm{e}{-5}$$

$$w_{ij} = \exp\left(\frac{1 - \tilde{d}_{ij}}{h}\right) \qquad h > 0 \text{ is a band-width parameter}$$

$$\mathrm{CX}_{ij} = w_{ij} / \sum_k w_{ik}$$

$$\mathrm{CX}(x, y) = \mathrm{CX}(X, Y) = \frac{1}{N} \sum_j \max_i \mathrm{CX}_{ij}$$

Given an image x and a target image y they represent each as a collection of points (e.g., VGG19 features): X = {xi} and Y = {yj}. They assume |Y | = |X|= N (and sample N points from the bigger set when |Y |!=|X|).

They extract the corresponding set of features from the images by passing them through a perceptual network Φ, where in all of their experiments Φ is VGG19. Let Φ l (x), Φ l (y) denote the feature maps extracted from layer l of the perceptual network Φ of the images x and y, respectively.

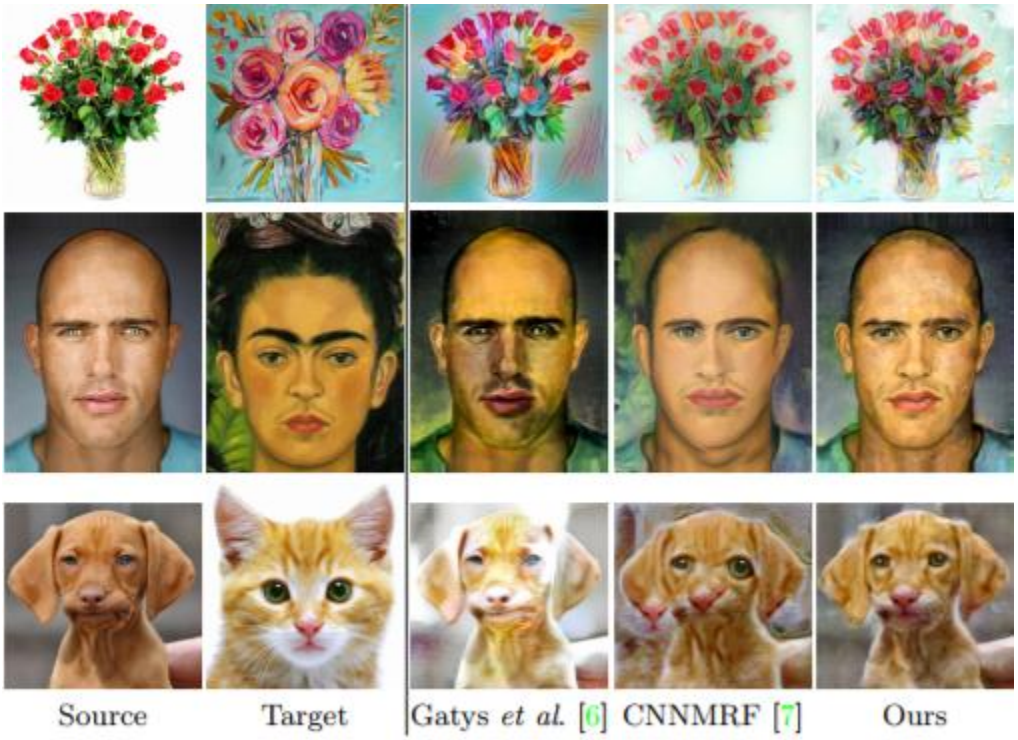
Other loss functions

- The Perceptual loss [8] $\mathcal{L}_P(x, y, l_P) = ||\Phi^{l_P}(x) - \Phi^{l_P}(y)||_1$, where Φ is VGG19 [31] and l_P represents the layer.
- The $L1$ loss $\mathcal{L}_1(x, y) = ||x - y||_1$.
- The $L2$ loss $\mathcal{L}_2(x, y) = ||x - y||_2$.
- The Gram loss [6] $\mathcal{L}_{Gram}(x, y, l_\mathcal{G}) = ||\mathcal{G}_\Phi^{l_\mathcal{G}}(x) - \mathcal{G}_\Phi^{l_\mathcal{G}}(y)||_F^2$, where the Gram matrices $\mathcal{G}_\Phi^{l_\mathcal{G}}$ of layer $l_\mathcal{G}$ of Φ are as defined in [6].

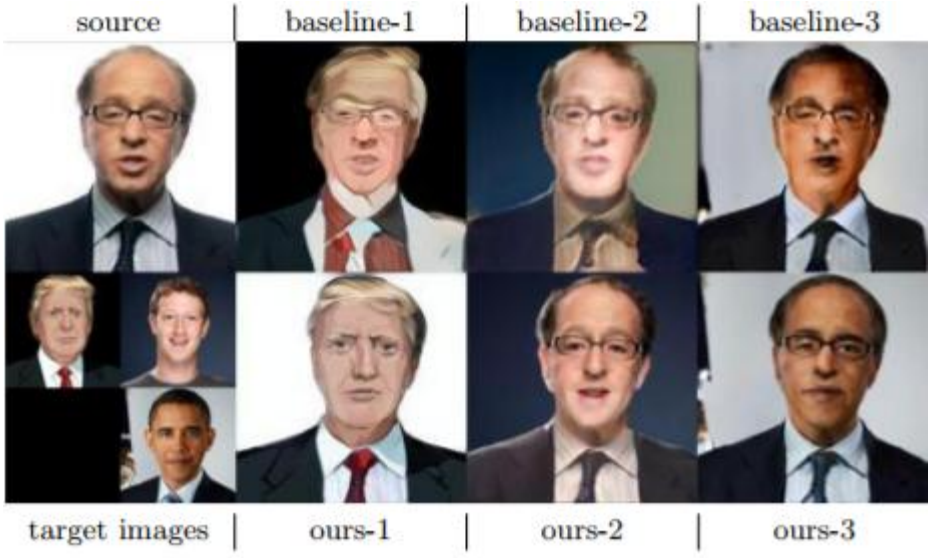
The first two are pixel-to-pixel loss functions that require alignment between the images x and y. The Gram loss is global and robust to pixel locations.

Applications

Semantic Style Transfer:



Single Image Animation:



Puppet control:

Unpaired domain transfer:

Discussion

Deformation is not obvious.