LEGAN: Disentangled Manipulation of Directional Lighting and Facial Expressions whilst Leveraging Human Perceptual Judgements

Sandipan Banerjee, Ajien Joshi, Prashant Mahajan*, Sneha Bhattacharya*, Survi Kyal, Taniya Mishra*

Affectiva Inc., Boston, USA

*Work done while at Affectiva

Overview

- Existing naturalness metrics either generate a single score for the whole dataset (FID) [1] or compute dissimilarity among image pairs (LPIPS) [2].
- Our quality metric rates the naturalness of individual synthetic face images in vacuum, serving as an auto substitute for human judgement.
- We directly plug this metric into LEGAN, our framework for disentangled lighting and expression manipulation, as an auxiliary discriminator.
- Using a set of hourglass nets, LEGAN separates the attribute sub-spaces & performs the desired translation while preserving identity.
- We build a quality estimation model (Q) to directly evaluate the perceived quality of GAN-generated images, and release the dataset of synthetic images along with their crowd-sourced quality annotations.
- When used in training, Q improves the perceptual quality of images synthesized by not only LEGAN but other off-the-shelf GANs as well.
- Q can also be used to filter face images synthesized by different models.
- LEGAN can be utilized as data augmenter to improve model performance on downstream tasks like face verification & expression recognition.

Contributions

- Dataset (URL): we collected face images generated using five different GAN & 3D model based synthesis approaches. After pre-processing, we ended up with 37K synth. images.
- Perceptual annotation: each image was scored for naturalness by 3 human raters. We used the mean (m) & standard deviation (std) from these ratings as the perceptual label.
- Quality estimation model (Q): as a cheap proxy for human annotation, we train a CNN with the images & their (m, std) labels. To capture the subjectiveness in visual perception, we formulated a margin based loss function for training.

Perceptual Quality Estimation

LEGAN: UTILIZING Q FOR LIGHTING & EXPRESSION MANIPULATION

- Architecture: LEGAN is composed of generator (G) and discriminator (D) networks, while Q serves as an auxiliary module for estimating quality of the synthesized images during training. Similar to other image-to-image translation models, LEGAN does not require paired data for training.

- Loss function: The full loss is a weighted sum of following:
  1. \( L_{adv} \): D’s weights are leveraged to tune G’s hallucinations to match distribution of real data and produce realistic samples as training progresses.
  2. \( L_{cls} \): ensures the target class association of a synthetic vector is preserved in the attribute space, using cross entropy over D’s softmax prediction.
  3. \( L_{rec} \): preserves structural integrity by cyclically reconstructing the input image from the translated output, comparing the two in pixel space.
  4. \( L_{wei} \): optimizes the perceptual quality of the translated output in the forward phase while preserving the same for the reconstructed input in the cyclic phase using Q’s prediction.

Experiential Results

- Training data: We use frontal face images from the MultiPIE dataset.
- Improving perceptual quality: Adding Q to the training framework improves visual quality and removes blob-like artifacts [4] from synthesized images (StarGAN: d).
- Improving off-the-shelf StarGAN: When added to the training framework of StarGAN [5], Q improves its performance on almost all metrics (compare rows 1 & 2).
- Correlation with existing metrics & human judgement: As can be seen in columns (2, 3, 6) & (6, 7) in the table above, our quality metric is well correlated with FID and LPIPS, and naturalness ratings provided by human annotators.

---