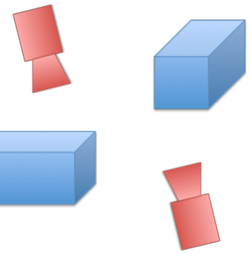
The Essence of Pose

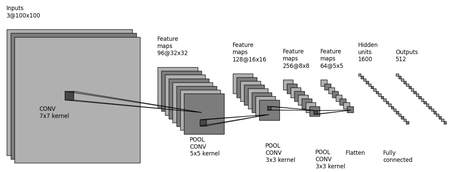
Jayant Thatte[1] Vincent Sitzmann[2], Timon Ruban[1]
Email: { jayantt }, { sitzmann }, { timon } @stanford.edu
Dept. of [1]Electrical Engineering and [2]Computer Science & Engineering, Stanford University

Motivation

Understanding 3D Vision



PoseNet



- From the Computational Vision and Geometry Lab[1]
- Does pose estimation and wide baseline matching

What does it learn?
What essential information does it preserve?

Example images from the dataset

References

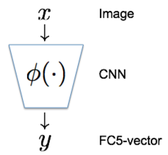
1. Computational Vision and Geometry Lab (CVGL) at Stanford, http://cvgl.stanford.edu/
2. Mahendran et al. “Understanding deep image representations by inverting them” CoRR Nov. 2014
3. J. Yosinski et al. “Understanding neural networks through deep visualization” CoRR Jun. 2015
4. Dosovitskiy et al. “Inverting Visual Representations with Convolutional Networks” CoRR Dec 2015
5. Zeiler et al. “Visualizing and Understanding Convolutional Networks” ECCV 2014
6. Simonyan et. al. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. ICLR 2014.
7. Google’s DeepDream, https://github.com/google/deepdream
8. Dosovitskiy et. al. “Generating Images with Perceptual Similarity Metrics based on Deep Networks” CoRR Feb 2016

Approach

Inverting FC5-Representation[2]

- Input: Vector of activations at specific layer of network
- Output: Reconstructed image through optimizing the loss function

$$\min_x \|\phi(x) - \phi(x_0)\|_2 + \lambda R(x)$$



- Hyper-Parameter Search has big influence on reconstructed images

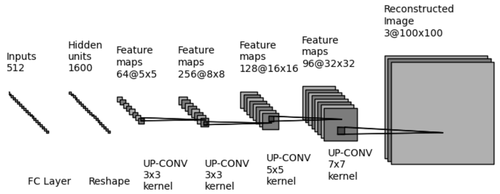
Relevant Priors and Regularizers

- Total Variation/Periodic blurring inhibits unnatural high-frequency patterns in reconstructed image
- L2 norm prevents maxing out magnitude of single pixels, ensuring realism and inhibiting overfitting
- Low-contribution clipping[3] linearizes the impact of single pixels on the loss function and sets pixels with little impact to zero.
- Generate image that maximizes the output of a single neuron

Additional Experiments:
- Inverting single neuron
- Saliency maps à la Simonyan[6]
- Deepdream à la Google[7]

Up-convolutional Generative Network (UGN)

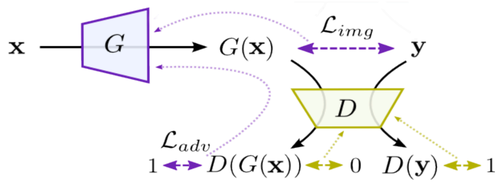
- Input: PoseNet vector embeddings
- Output: Reconstructed images through repeated up-convolutions[4, 5]
- Data: Trained on ~6 million images
- Training: MSE reconstruction loss
  Batch normalization at each layer
  Update rule: Adam (more robust)
- Random Hyper-parameter search over 500 combinations sampled from a logarithmic search space



UGN Architecture

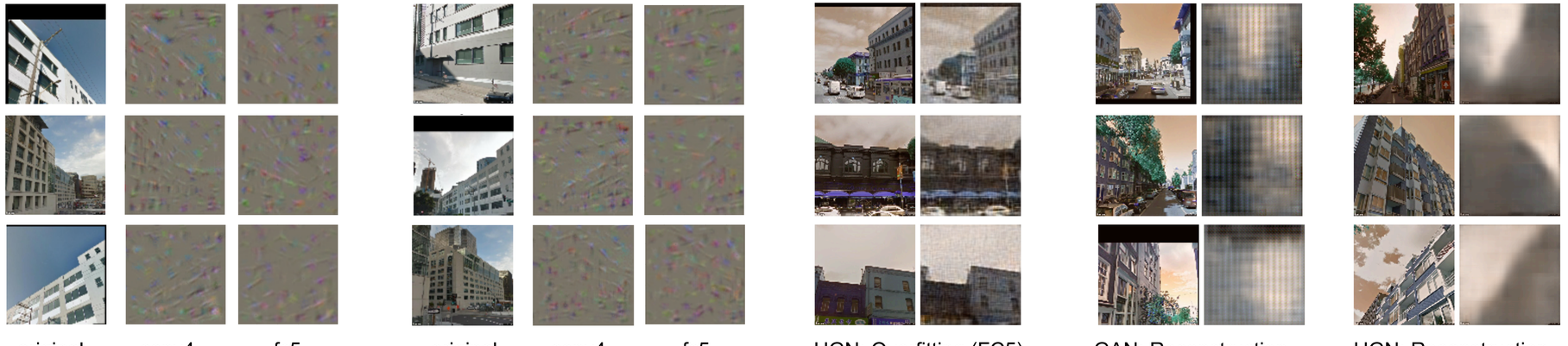
Generative Adversarial Network (GAN)

- L2 reconstruction loss sub-optimal[8]
- Discriminator: Learns to tell generated images from groundtruth images
- Generator: Learns to produce images Good enough to fool the discriminator
- Discriminator: Dropout, no batchnorm
- Generator: Same architecture as UGN



GAN high-level block diagram [8]

Experimental Results



original conv4 fc5 original conv4 fc5 UGN: Overfitting (FC5) GAN: Reconstruction UGN: Reconstruction