Image Mosaic

Gonçalo Gil

CS 229 Final Project, Stanford University, gilg@stanford.edu

I Introduction

In this project we classify a set of art photographs by fine art photographer Mário Cabrita Gil. We employ Convolutional Neural Networks with the VGG16 and VGG19 architecture to classify photographs that span four decades of professional photographic work. For each photograph, the 4096 dimension feature vector from the last convolutional layer is reduced to two-dimensions by employing distributed Stochastic Neighbor Embedding (t-SNE). Finally, the photographs are arranged in a 2D mosaic according to their similarity.

I. Image recognition (VGG16/19)

The model is trained on a subset of the ImageNet database, and was employed in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). The VGG16/19 network is trained on more than a million images and can classify images into 1000 object categories.

II.Dimensionality reduction (PCA, t-SNE)

To reduce the feature vector to two components we employ Principal Component Analysis (PCA) and a variation to the Stochastic Neighbor Embedding technique called t-SNE [Maaten and Hinton, 2008].

III. Mosaic (RasterFairy)

The 2D image cloud is transformed into a mosaic with an arbitrary shape, while preserving neighborhood relations present in the original set. We use the RasterFairy python library developed by Mario Klingemann which is related to Isomatch [O. Fried et al., 2015] and Kernelized Sorting [N. Quadrianto et al., 2009].

II Methods

CNN

The input to the ConvNet is a fixed size 224×224 Red-Green-Blue (RGB) image. Preprocessing involves subtracting the mean training set RGB value from each pixel. The ConvNet configuration has a smaller receptive field in the first convolution layer than previously employed. Namely, in the VGG16 architecture, the receptive field is 3×3 and the convolution stride is 1 pixel. There are two fully connected layers with 4096 units each and ReLU activation. In the first layer, the network learns 64 filters with size 3×3 along the input depth with a bias for each filter. Hence, the number of parameters in the first convolutional layer is 42×3×3×3+64=1792 parameters. The number of filters increases by a factor of 2 after each max-pooling layer, until it reaches 512. The total number of parameters for VGG16 is 138 million parameters.

t-SNE

Given a set of N high-dimensional objects $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, t-SNE first computes probabilities p_{ij} that are proportional to the similarity of objects $\mathbf{x}_i$ and $\mathbf{x}_j$,

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i^2)}. \quad (1)$$

Next, it computes the similarities

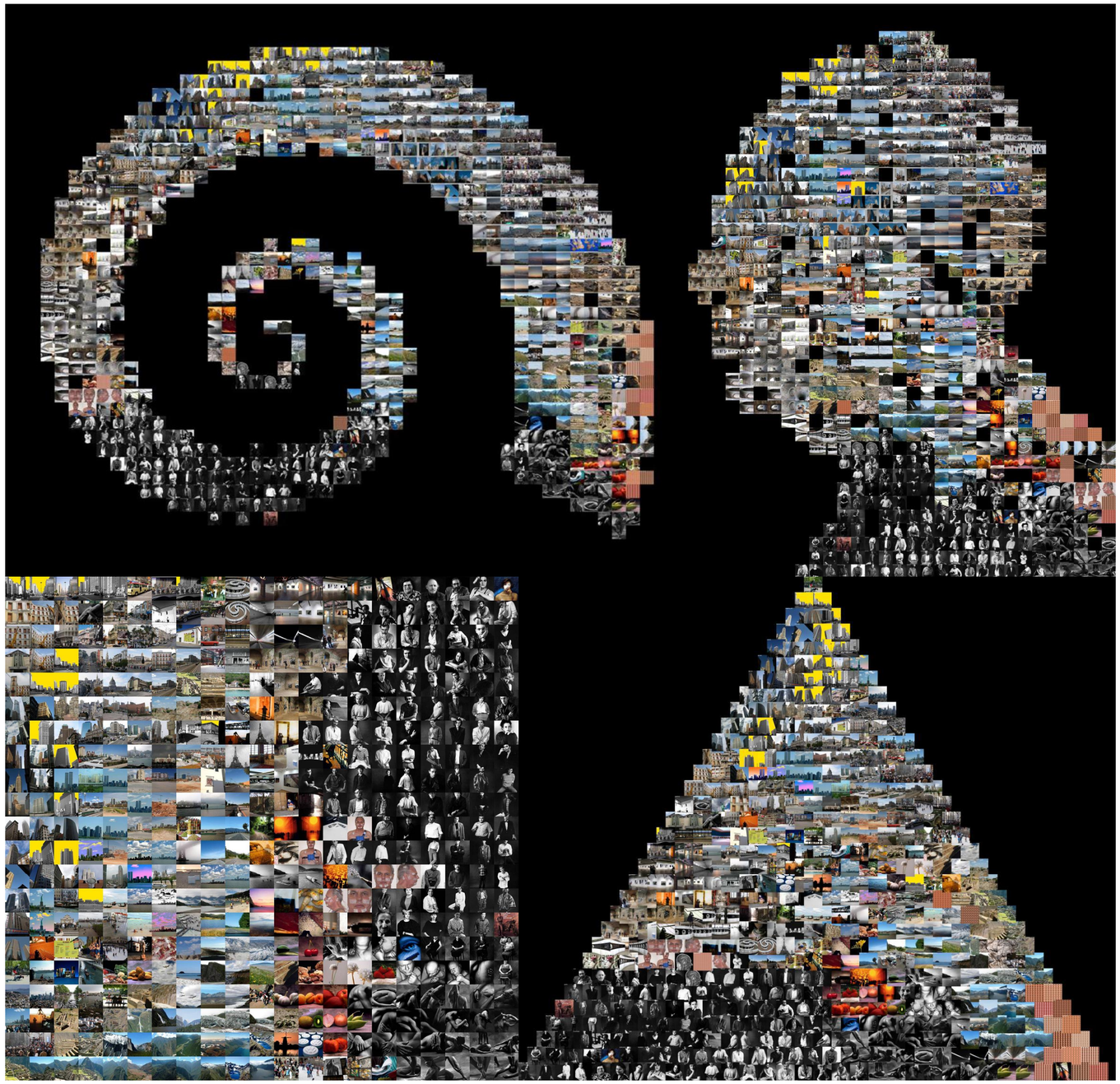
$$q_{ij} = \frac{(1+||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k\neq i}(1+||\mathbf{y}_k - \mathbf{y}_m||^2)^{-1}} \quad (2)$$

between low-dimensional points to learn a **d**-dimensional map, where $\{\mathbf{y}_1, ..., \mathbf{y}_N\}$, where $\mathbf{y}_i \in \mathbb{R}^d$. Finally, the $\mathbf{y}_i$ are computed by minimizing the Kullback-Leibler (KL) divergence of the distribution **Q** from the distribution **P** using gradient descent,

$$KL(P||Q) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}}. \quad (3)$$

The optimization problem results in a map that reflects the similarities between the high-dimensional inputs.

III Results



IV Discussion

The idea behind this work was to present a visually pleasing mosaic of a set of photographs. We attempted different architectures for the classification task, such as VGG16 and VGG19 and found that, as far as the visual outcome, the results are very similar. There are some issues with the image classification, but overall the results are excellent. Shown above is the t-SNE version of the dimensionality reduction step, which was slightly better than that found with PCA. One of the parameters that can be used to tune t-SNE is perplexity. We found that larger perplexity values do better than smaller values, but overall the difference is small. The mosaic shape can be chosen arbitrarily to produce visually pleasing presentations. In the future, we would like to use a clustering model to separate similar groups of images into their own "islands" in the mosaic.

V References

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning.Nature, 521(7553):436-444, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211-252, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems pages 1097–1105, 2012.

Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In Advances in neural information processing systems, pages 857-864, 2003.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(Nov):2579–2605, 2008.

VI Acknowledgements

Images by Mário Cabrita Gil
http://www.mariocabritagil.com

RasterFairy python Library by Mario Klingemann
https://github.com/Quasimondo/RasterFairy