Image Mosaic

Gonçalo Gil

Stanford University
gilg@stanford.edu

December 15, 2017

Abstract

We employ Convolutional Neural Networks with the VGG16 and VGG19 architecture to classify photographs that span four decades of professional photographic work by Mário Cabrita Gil. For each photograph, the 4096 *dimension feature vector is reduced to two-dimensions by employing distributed Stochast Neighbor Embedding (t-SNE). Finally, the photographs are arranged according to their similarity value in a mosaic display.*

I. Introduction

With the advent of large-scale distributed computer clusters and increased computational power, it has become feasible to employ Convolutional Neural Networks (ConvNets) for a wide range of applications. Efficient use of Graphics Processing Units (GPUs), rectifiers (ReLU) and dropout regularization, has catalyzed a revolution in computer vision applications. ConvNets have been successful in identifying faces, objects and natural settings and is widely used in robotic vision and self-driving vehicles. Other applications of ConvNets include natural language understanding and speech recognition [LeCun et al., 2015].

ConvNets are a multi-layer neural network trained on a back-propagation algorithm with a special architecture designed to recognize visual patterns directly from pixels with very little preprocessing. Nearly 3 decades ago, LeCun et al. [1989] was the first paper on convolutional networks trained by backpropagation with the task of classifying low-resolution images of handwritten digits.

In this project we will classify a set of art photographs by author Mário Cabrita Gil [Mário Cabrita Gil, 2017] by employing ConvNets and visualize the results in a 2D mosaic arrangement according to similarities between images.

II. Dataset and Features

The dataset consists of 725 images in total and most are available at the author's website [Mário Cabrita Gil, 2017]. There are two preprocessing steps before the images are fed into the ConvNet. First, each image is rescaled to 224×224 because the ConvNet takes as input RGB images of fixed size. Second, to help increase speed and accuracy, we subtract the mean value of the RGB values for each pixel. We employed both VGG16 and VGG19 architectures. One of the images, depicted in Figure 1, is an aerial photograph of the New Zealand Alps. For this image the classification network has made the predictions listed in Table 1. The result is surprisingly good as it correctly identified it as an Alp.

Label	Probability
Alp	0.902
Valley	0.030
Cliff	0.026
Volcano	0.010
Ibex	0.006

Table 1: *Classification of image in Figure 1.*

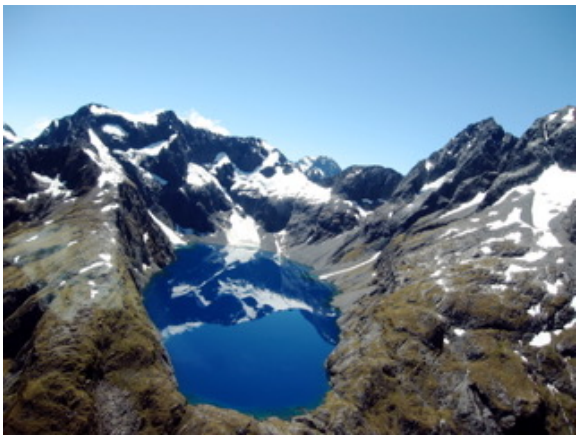

Figure 1: *A picture of the New Zealand Alps*

In the dimensionality reduction step we employed both PCA and t-SNE. In the former case, there are no tunable parameters except for the number of dimensions. However, in the latter case, there is an important and somewhat mysterious parameter called perplexity, p. Perplexity is a measure of how the algorithm manages focus between local and global aspects of the data [Wattenberg et al., 2016]. The results are not necessarily straightforward, hence we ran the model with a range of perplexity values between 10 and 100.

For the mosaic, several arrangements were attempted in order to generate a pleasing visual. We attempted several geometrical shapes, including, spiral, circle, rectangle, triangle, as well as arbitrarily shaped monochrome images, such as the artist's initials, side profile and country of birth.

III. Methods

i. Image recognition

To classify the images we employ the pretrained VGG16 ConvNet model. The model is trained on a subset of the ImageNet database ImageNet [2017], and was employed in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [Russakovsky et al., 2015]. VGG-16 is trained on more than a million images and can classify images into 1000 object categories and has learned rich feature representations for a wide range of images. The main breakthrough is that VGG16 it increases the depth of the weight layers to 16 layers.

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, comparing to a 11×11 receptive field with stride 4 [Krizhevsky et al., 2012], or 7×7 receptive field with stride 2 [Sermanet et al., 2013, Zeiler and Fergus, 2014]). The max pooling is 2×2 with a stride of 2 pixels.

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 $N_p^{[1]} = 42 \times 3 \times 3 \times 3 + 64 = 1792$ parameters. The number of filters increases by a factor of 2 after each max-pooling layer, until it reaches 512. Following a similar procedure as above, the total number of parameters for VGG16 (referred to as network D in Simonyan and Zisserman [2014]) is 138 million parameters. The output layer employs softmax classification and has 1000 units which represent the 1000 ImageNet classes [Simonyan and Zisserman, 2014]. The previous layer before classification has 4096 units and that is the feature vector used for dimensionality reduction because it produces better results.

ii. Dimensionality reduction

Several dimensionality reductions techniques for visualizing high-dimensional data exist, such as Principal Component Analysis (PCA). For high-dimensional sets of nonlinear nature, kernel PCA can be used to address the problem of nonlinear dimensionality. Several algo-

rithms can be cast as kernel PCA, among which are Local Linear Embedding, Metric Multidimensional Scaling, Laplacian Eigenmaps and Isomap. In this paper, we employ the a variation to the Stochastic Neighbor Embedding technique [Hinton and Roweis, 2003] called t-SNE [Maaten and Hinton, 2008]. The t-SNE algorithm is easier to optimize and leads to significantly improved visualizations when compared to the original technique.

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

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

between low-dimensional points to learn a $\mathbf{d}$-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 $\mathbf{Q}$ from the distribution $\mathbf{P}$ 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.

IV. Results

The idea behind this work was to present a visually pleasing mosaic of a set of art 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. We attempted to run the dimensionality reduction algorithm on the output layer (1000 dimension vector) and the previous layer (4096 dimension vector) and found that the latter is much better than the former. The tuning parameters and dimensionality reduction algorithms attempted do not change the outcome significantly as far as the final visual impact. Art is a subjective matter and it is difficult to say what is an optimum outcome as different people perceive art in a very personal way. In that sense we find very little difference in applying the VGG16 or VGG19 architectures, and as far as dimensionality reduction, PCA and t-SNE produced subjectively similar results. This is shown in Figure 2 with the final mosaic in a rectangular shape with the VGG16 and VGG19 architectures and t-SNE perplexity $p = 15, 30, 45, 60$.

To highlight the difference between perplexity values, we show the nearest neighbors of an image of a cherry for two different perplexity values in Figure 3. In one case, where $p = 45$, the model correctly identifies the cherry as belonging to a group of similar images of fruits or food and a round object that resembles an orange fruit. However, by changing the perplexity value by a small amount ($p = 35$) the cherry is now placed near seemingly unrelated objects, such as set of perfumes, a fountain, a beach, buildings and an old car. Interestingly, in both cases it considered the yellow flower as a nearest-neighbor.

Finally, in additional to rectangular shapes, the mosaic shape can be chosen arbitrarily to produce visually pleasing presentations. In Figure 4 we show some of the different arrangements attempted, including the side profile of the artist, a triangle and a spiral. Again, any attempt at quantifying the results is subjective. To the author of this paper, the most appealing result is the spiral. However, that conclusion may very well change from observer to observer.

V. Conclusion

For this project, we have built a pipeline where a set of images is classified with a ConvNet and the extracted feature vector is reduced to two-dimensions using PCA and t-SNE. Then,

Figure 2: *Image showing the results for different values of perplexity for the t-SNE dimensionality reduction algorithm. The VGG16 architecture is on the left, while the VGG19 architecture is on the right. The values of perplexity from top to bottom are* $p = 15, 30, 45, 60$.

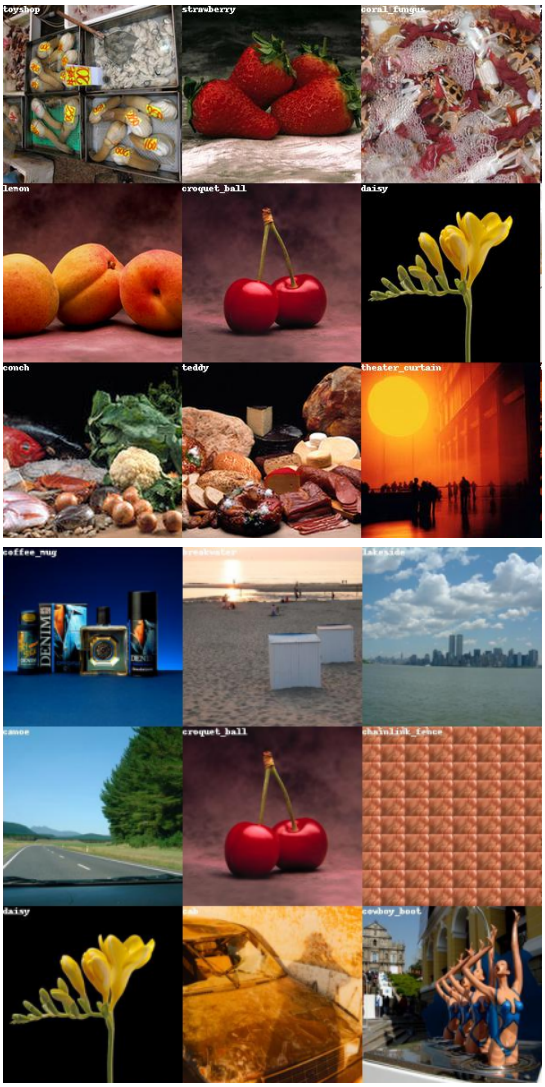

Figure 3: *Nearest neighbors of a cherry for two values of perplexity,* p. *The top image shows a good arrangement for the cherry whereas the bottom image shows a poor arrangement for the cherry. Both use VGG16 with* $p = 45$ *for the top image and* $p = 35$ *for the bottom image.*

we generate a mosaic of arbitrary shape with the entire set of photographs arranged by similarity between each photograph. We also attempted to use auto-encoding as another way to reduce dimensions but unfortunately it did not yield good results because of computer power limitations since the larges image size that could be processed on our machines was limited to 60×60 pixels. In the future, we would like to use a clustering model to separate similar groups of images into their own separate islands in the mosaic.

VI. Contributions

All work by Gonçalo Gil.

VII. Acknowledgments

We thank Mário Cabrita Gil for the images and Mario Klingemann for the RasterFairy python library used to generate the mosaics of arbitrary shape.

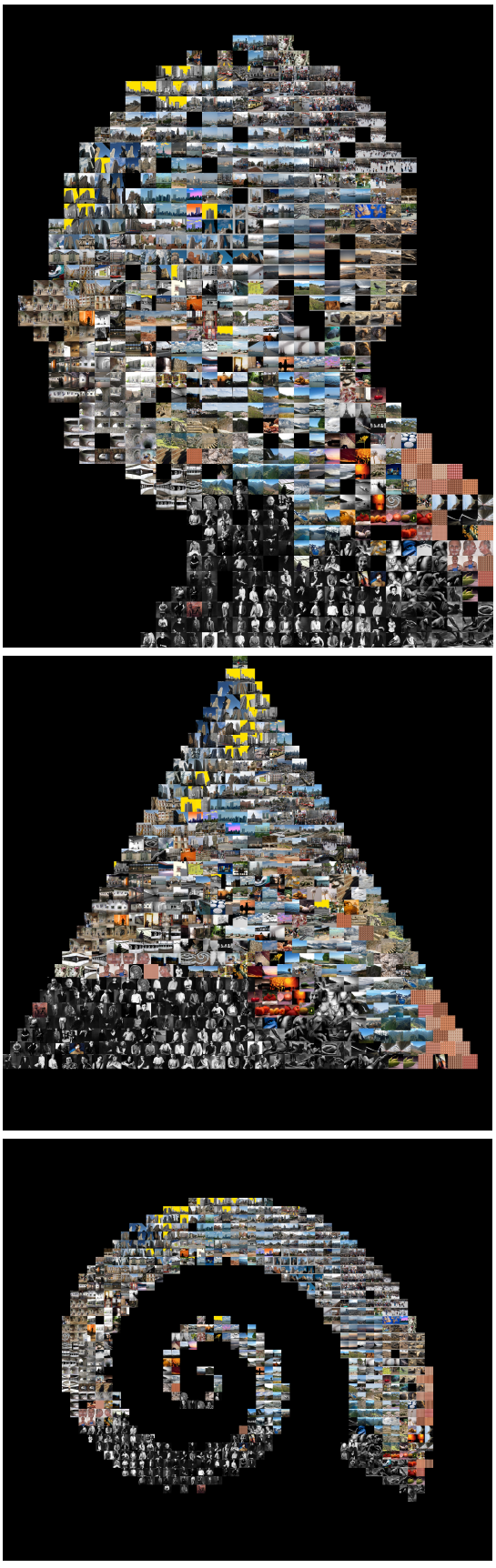

Figure 4: *The mosaic with different shapes.*

References

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

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551, 1989.

Mário Cabrita Gil. Mário Cabrita Gil, 2017. URL `http://www.mariocabritagil.com/`. [Online; accessed 20-November-2017].

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.

ImageNet. Imagenet, 2017. URL `http://www.image-net.org`. [Online; accessed 20-November-2017].

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.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint 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.