# Explainability pipeline for urban images


## ABSTRACT

Recent advances in deep neural networks and generative adversarial networks allow us to both create discriminative models and invert the processes to enable generative models, capable of generating image samples very close to real world natural images. This has proven a useful tool in visualizing what a network learns when it learns to classify cats, dogs and objects. The challenge is to understand whether the same approaches can be useful in understanding much more abstract and meta properties like Beauty, Liveliness, depression etc. This paper proposes a pipeline and the necessary framework to infer affective components in urban images.



**ACM Reference format:**
. 2017. Explainability pipeline for urban images. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 3 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn


## 1 INTRODUCTION

Deep neural nets are progressing at an amazing pace over the past decade. The community as a whole has been breaking new ceilings when it comes to classification and inference records for specific tasks like object detection, scene detection, language modeling etc. But the internal workings and the internal process of neural nets before coming to a particular decision, more or less still remains a mystery. Neural nets have more or less remained a black box for its users. Explain-ability and understanding the deep reasoning behind decisions is one of the most researched problems in the machine learning community.

The problem of explain-ability becomes even more abstract and obscured, when we are dealing with tasks that handle meta, and abstract quantities like sentiment, affects and aesthetics. Despite the black box like nature, deep neural networks have done remarkable strides in understanding creativity [10], memorability [6] or beauty [12]. These works explore perceptual qualities of media objects using deep learning, and treat the explainability of the models using round about methods like perturbation of input and understanding correlation of several governing variables with decisions of the network etc. These methods are perfectly valid and do give some interesting insights into the decision influencing factors for the models, however still fail to explain the decision making process of the model itself.

This paper builds on top of several works done before, in the areas of explainability of neural nets and understanding affective dimensions as listed before takes a step towards extending these to the realm of urban emotions and aesthetics. More so we strive to propose a generalizable pipeline for analyzing geo-referenced images.



*Conference'17, Washington, DC, USA*



DOI: 10.1145/nnnnnnn.nnnnnnn


## 2 PIPELINE

With the motivation of creating a streamlined framework, we propose an end-to-end pipeline for explaining urban image categories, which is illustrated in Figure 1.The system allows anyone with an arbitrarily set of annotated image data $X = I_1, I_2....I_n$ annotated in classes $Y = y_1, y_2, ..., y_k$, to transform natural images between classes: the pipeline can transform an arbitrary image $I_i$ belonging to class $y_i \in Y$ , to image $I_j$ from class $y_j \in Y$. This allows to visually reason about the discriminative properties between classes $y_i, y_j \in Y$, and visually understand what are the salient characteristics that make it to classify images into the respective classes $y_i, y_j$. These questions might be trivial for tangible classes of objects, but still remain largely unexplored for intangible classes representing concepts like affects and aesthetics.

This pipeline approaches the transformation problem in two phases. Assuming the previous example where we transform images from class $y_i$ to class $y_j$, the first step is to produce a prototype image or a template image $\hat{I}_j$, that represents the basic traits of the the destination class $y_j \in Y$. The second step is to match this template image $\hat{I}_j$, with the closest natural image In mathematical terms, we want to choose a different image $I'$ from $X$ so as to minimize $E(I', \hat{I}_j)$ , where $E(I_1, I_2)$ is some error measure that quantifies visual error between two images. This image $I'$ is effectively a natural transformed image.

### 2.1 Data

We assume that the input to the framework is simply a crawl of geo-referenced images, annotated for an arbitrary use case. The data needs to be preprocessed in two separate ways. Data clustering and Data Augmentation

*2.1.1 Data Clustering.* In the first step, we cluster data on the basis of some pre-determined attributes. The main aim of data clustering is to reduce the diversity and variance in the data. The current state of the art deep-learning models work on highly specific classes of objects with low variance in the image semantics. When it comes to urban images, the variance in structure and composition between two images is extremely high. Some clustering methods are listed below

- The most simple yet effective way to cluster data could be based on geographical context. This can be done simply by using geographical boundaries of areas of interest and clustering images based on attributes like rural, urban, sub-urban, city etc. . This seems like an intuitive pre-processing step, as images from countryside look widely different compared to the images from the urban environment, and might be very diverse despite having similar annotations.

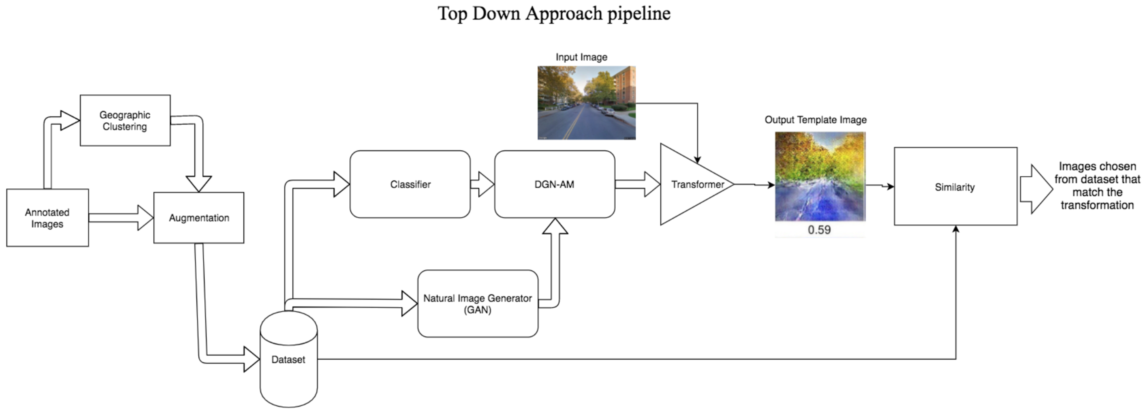


**Figure 1: Process pipeline for the "Top Down" approach for explainability**

- You may even cluster images based on some visual/latent similarity measure. One possible way to do this is by extracting higher dimensional features, and clustering images based on these features.

#### 2.1.2 *Data Augmentation.*

In the due process of clustering, it is expected that the total size of data available to train would reduce. Smaller data size implies that a machine learning model has a risk of over-fitting and in the worst case not learning anything at all. Hence there is a need to augment this clustered data with some additional real and transformed data. Some of the techniques that may be used are listed below.

- The most common augmentation technique in deep learning literature is to do transformations on the image. Transformations like flipping images, cropping, adding noise , shifting color histograms can increase the data points for training and at the same time reduce the risks of over fitting.
- Because the images are geo tagged, one can augment the data by acquiring additional images which fall very close geographically to the original image. In this approach, care must be taken to maintain visual similarity of additional images. This can be achieved by several ways including, but not limited to, using higher dimensional features extracted using some pre-trained image models, to measure visual similarity (as described in the clustering section).

.

## 2.2 Template Generator

To produce the template image $\hat{I_j}$, we need two components in place, 1) A classifier which learns how to distinguish between different image categories 2) A generative model (GAN), that can generate samples from the distribution of the dataset images. 3) An activation maximization framework, that, based on the GAN generator, generates images that maximizes activation for a given annotation class [2] (our template images).

- Classifier. In order to produce template images, we now train a deep learning based classifier, that, given an image $I$, can correctly classify it in one of the $k$ classes: the aim of this pipeline is to explain *what* the classifier is learning about the annotations. The assumption here is, once the classifier learns to discriminate amongst the classes, it also learns discriminative properties about the images that fall in those annotation categories. Several works have shown that as a convolutional network trains, several semantic and object detectors emerge at higher abstraction levels in the network architecture.
- *Generator*. We train a generative adversarial network (GAN) which can generate an approximate natural looking image drawn from distribution of a particular class of images. For our use case we train the GAN mentioned in [2]. This GAN generator would learn to generate a natural-like image that represents the overall structure and knowledge about the Dataset.
- *Activation Maximisation*. We plug in the GAN and the classifier networks into the Activation maximization network so as to generate an ideal image that maximizes the activation for a particular class. The output of this step would be an image, which is very close to a natural image that the classifier is trained on, but has all the right motifs that maximizes the annotation class. Essentially this image is a representation of the overall knowledge about a particular annotation class, that the classifier network has learned through training.

For an image I We call this image as **template image**. So in a sense this step is transforming a natural image into an image that maximizes an annotation class.

## 2.3 Similarity and inference

In this final step we find images from the dataset that are closely aligned, in terms of some visual similarity metric $E(I_1, I_2)$, with the generated template image. The result of this exercise is to find the most similar looking image to an input image that maximizes a particular annotation class. So in a sense we are transforming one natural image into another natural image so as to maximize some annotation class. The visual differences in these two natural images, can act as the subject of reasoning for the explainability.

## REFERENCES


[1] Ritendra Datta and others. 2008. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *2008 15th IEEE ICIP*. IEEE, 105–108.

[2] Alexey Dosovitskiy and Thomas Brox. 2016. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4829–4837.

[3] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. *arXiv preprint arXiv:1608.01769* (2016).

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[5] Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkillâĎć: a Bayesian skill rating system. In *Advances in neural information processing systems*. 569–576.

[6] Phillip Isola and others. 2011. What makes an image memorable?. In *IEEE CVPR*. 145–152.

[7] Aditya Khosla and others. 2014. What makes an image popular?. In *Proceedings of the 23rd WWW*. International World Wide Web Conferences Steering Committee, 867–876.

[8] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. 2014. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 779–785.

[9] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*. 3387–3395.

[10] Miriam Redi and Others. 2014. 6 seconds of sound and vision: Creativity in micro-videos. In *Proceedings of the IEEE CVPR*. 4272–4279.

[11] Philip Salesses, Katja Schechtner, and César A Hidalgo. 2013. The collaborative image of the city: mapping the inequality of urban perception. *PloS one* 8, 7 (2013), e68400.

[12] Rossano Schifanella and others. 2015. An Image is Worth More than a Thousand Favorites: Surfacing the Hidden Beauty of Flickr Pictures. In *Proceedings of THE 9TH ICWSM 2015*.

[13] Yilin Wang and others. 2015. Unsupervised Sentiment Analysis for Social Media Images. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press. http://dl.acm.org/citation.cfm?id=2832415.2832579

[14] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.