CPM2 – Saliency Detection

Jarosław Biedrzycki

1. Introduction

Artificial neural networks have been proven to be useful in computational modeling of cognitive processes. For example, convolutional neural networks, because of their similarity to human primary visual cortex, have been very effective in solving problems in the field of computer vision. One such problem is saliency detection, which is detecting the image areas that grab people's attention most.

The saliency detection is a complex process that involves different subprocesses. There might be different reasons why certain parts of an image stand out. Some low-level features of the visual stimuli are more easily detected than others. This is also true for high-level features like faces. Because of their hierarchical nature, the deep neural networks might be able to integrate features at different levels to predict the most salient areas.

This project is an attempt to build a deep neural network that would be capable of predicting the most salient areas in images of different kind.

2. Materials and Methods

2.1. Data set

The project was based on a publicly available subset of CAT2000 data set, which was designed for saliency research (Borji and Itti, 2015). It consists of 2000 images from 20 categories (see Figure 3 for the list) along with corresponding fixation maps acquired from 18 participants during 5 seconds of free viewing for each image. Some of the categories strongly elicit bottom-up attentional responses, while others significantly elicit top-down factors. Therefore the data set can be used to study different aspects of visual attention.

The data was split into two subsets: 75% of the observations was be used for training, while the remaining 20% constituted a test set for the models. The observations were assigned to each subset randomly. Special care was taken to ensure that the categories were represented with an equal number of items. This resulted in 1500 observations (75 per category) in the training set and 500 (25 per category) in the test set.

2.2. Models

Ground Truth

The saliency maps referred as Ground Truth were generated by applying a Gaussian filter (sigma = 40) to the fixation maps obtained during the eye-tracker procedure. Ground Truth acts as the best case scenario.

Blind Truth

Blind Truth utilizes a constant saliency map instead of creating an individual one for each stimulus (see Figure 1). The map is characterized by a blurred blob in the middle of the screen. The term *blind truth* came up as a joke after realizing that this kind of saliency map yield better results than any model deliberately constructed by the author.

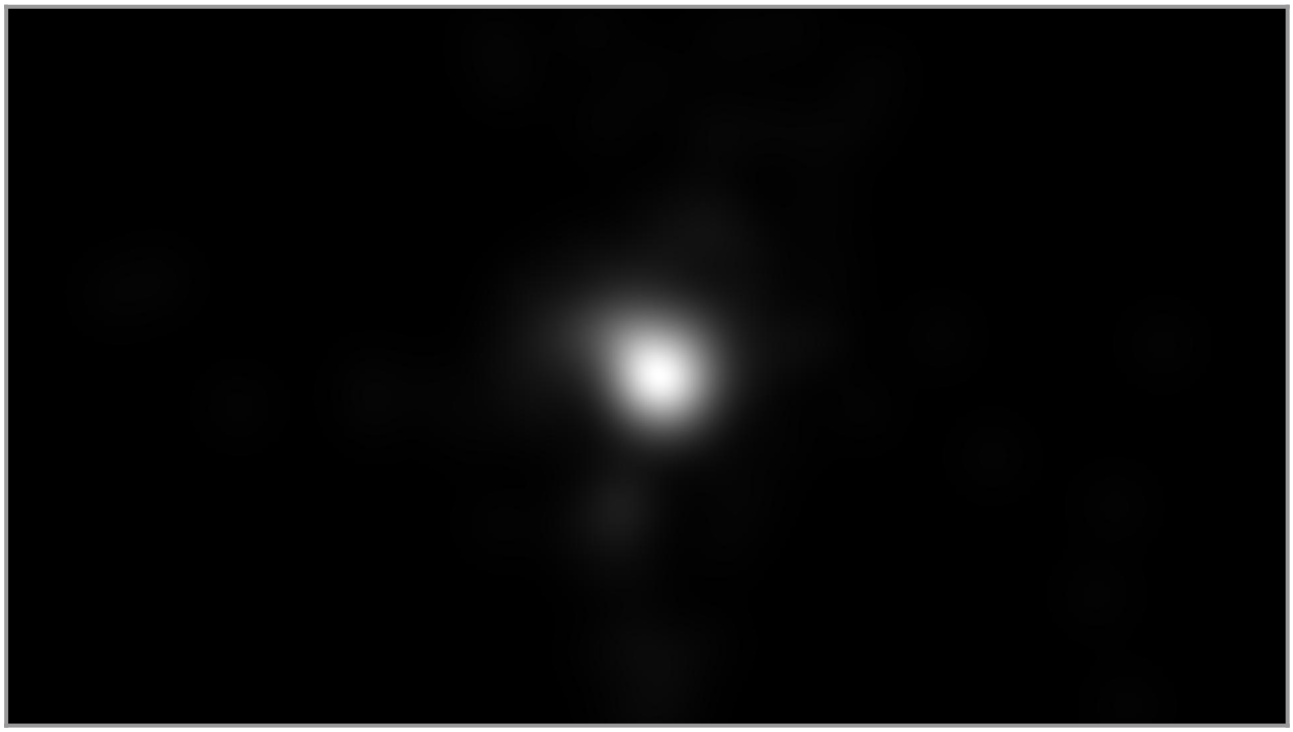


Figure 1: The saliency map used as Blind Truth

Fully Convolutional Network

The Fully Convolutional Network (FCN) built in this project predicts saliency maps that correspond to images provided in the input. The network consists of two parts. The first one is based on a regular VGG16 model, pre-trained on ImageNet data set. Multiple blocks (1-5) are connected to create a sequence of convolutional and max pooling layers. Different number of filters are used depending on the block. The data is downsampled by a factor of 2 at the end of each block. As the image is processed further by the following layers, the representations become more abstract. The second part of the network

constructs a saliency map from those abstract representations. This was achieved by using blocks of deconvolutional and upsampling layers. In contrast to the VGG16, the blocks (6-10) are not linked sequentially. Instead, they combine multiple outputs from previous layers in order to enhance the spatial precision of the constructed saliency map. This is illustrated best in Figure 2. All layers in the network apply a ReLU activation function. The only exception is the last layer (block 10) which uses a sigmoid function.

The further configuration of the network is as follows. The VGG16 layers were frozen in order to prevent changing of their weights. Binary cross-entropy was used as the loss function, while the ADADELTA algorithm was applied as the optimizer. The network was trained using the data from the training subset. Learning was performed in 10 epochs, with a batch size of 10. All calculations were performed using CRAN R with Keras and Tensorflow packages.

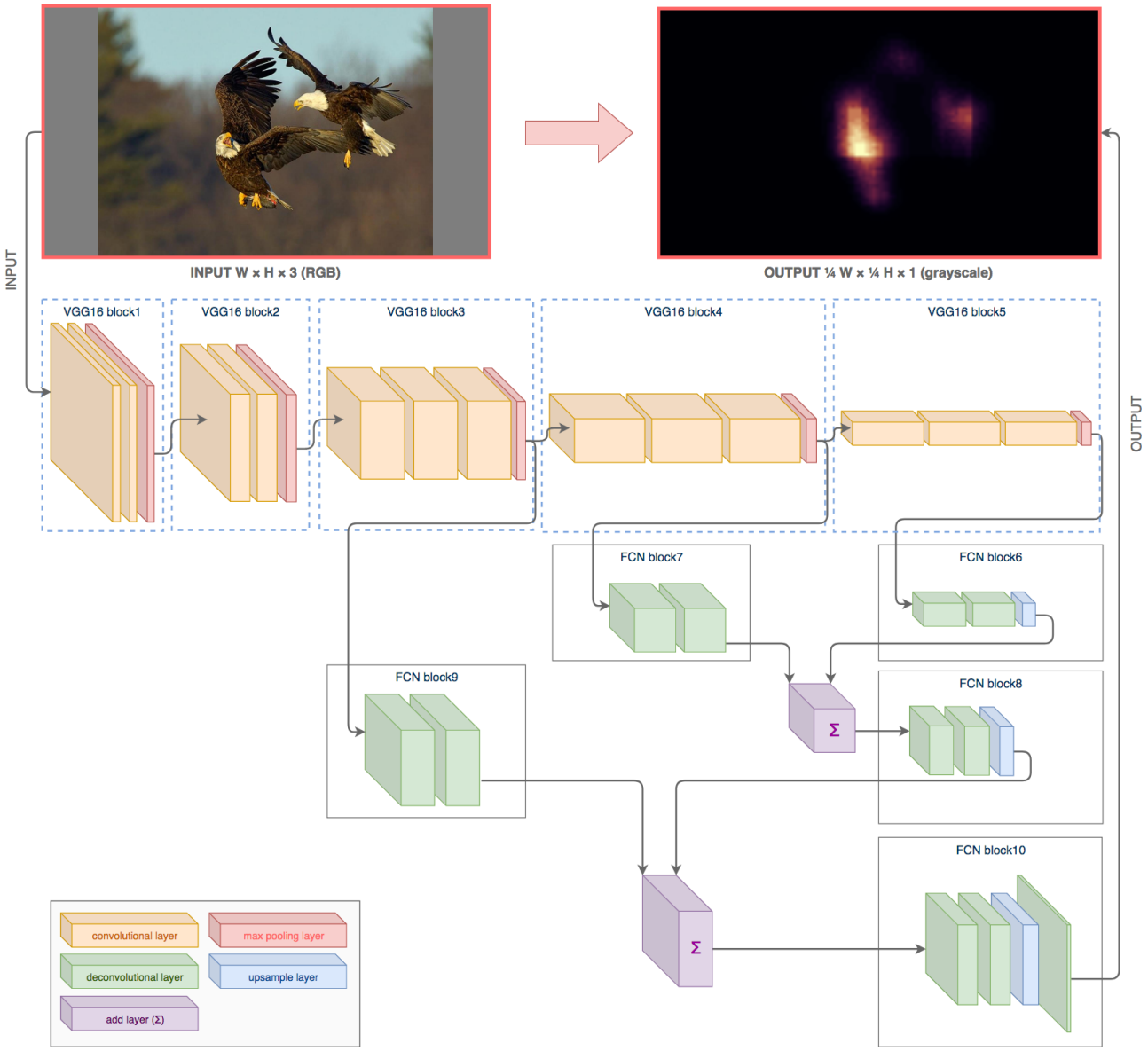


Figure 2: FCN network architecture. Top) The model maps images into predicted saliency maps . Bottom) The blue dashed lines indicate VGG16 blocks, while the gray solid ones correspond to the constructive blocks of the network.

2.3. Prediction and Model Accuracy

The aforementioned models were used to generate saliency maps for all stimuli in the test subset. In order to quantify the correspondence between the true fixations and the saliency maps, Normalized Scanpath Saliency (NSS) method was used.

"Given a saliency map P and a binary map of fixation locations Q^B:

$$NSS(P, Q^B) = \frac{1}{N} \sum_i \overline{P_i} \times Q_i^B$$

$$\text{where } N = \sum_i Q_i^B \text{ and } \overline{P} = \frac{P - \mu(P)}{\sigma(P)},$$

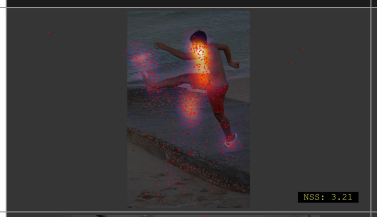
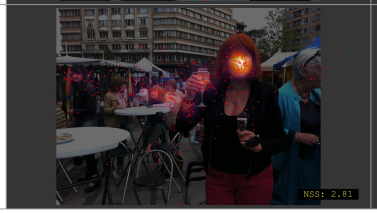
where i indexes the i-th pixel, and N is the total number of fixated pixels. Chance is at 0, positive NSS indicates correspondence between maps above chance, and negative NSS indicates anti-correspondence." (Bylinski et al., 2019)

2.4. Statistical Analysis

For each category of stimuli a paired t-Student test was performed to compare the NSS values between the Blind Truth and the FCN model. Bonferroni correction for multiple comparisons was used therefore the significance level was set to α = 0.0025.

3. Results

The saliency maps were generated using all 3 models (Ground Truth, Blind Truth and FCN). Some examples for selected categories are presented in Table 1.

Category	Ground Truth	Blind Truth	FCN Prediction
Action			
Indoor			
Affective			
Social			

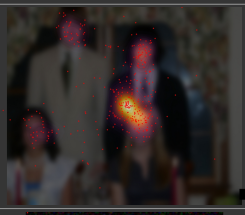
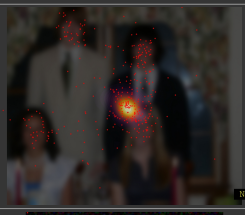
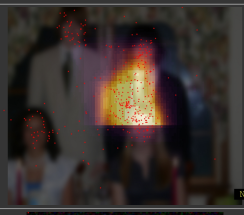
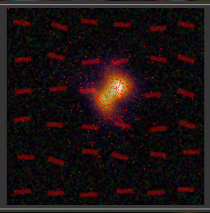
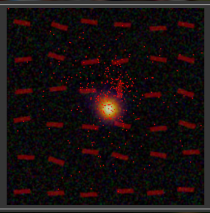
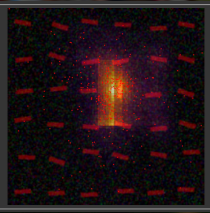
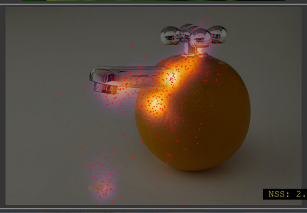
Category	Ground Truth	Blind Truth	FCN Prediction
Outdoor Natural			
Low Resolution			
Pattern			
Object			
Art			
Outdoor Man Made			

Table 1: Selected stimuli overlaid with saliency maps corresponding to different computational models (ground truth, blind truth and FCN prediction). The fixations registered during the eye-tracking procedure were added to the images in order to allow better comparison between the model predictions and registered data. NSS calculated for each prediction is available in the bottom right corner.

The statistical comparison of the NSS values revealed the following significant differences: In Action ($p < 0.0001$) and Sketch ($p < 0.0001$) categories the NSS values were higher for FCN predictions compared to Blind Truth. The opposite pattern (Blind Truth > FCN) was observed in LowResolution ($p < 0.0001$), OutdoorNatural ($p < 0.0001$) and Satellite ($p <$

0.0001) categories. Mean NSS values for different categories of stimuli can be found in Figure 3.

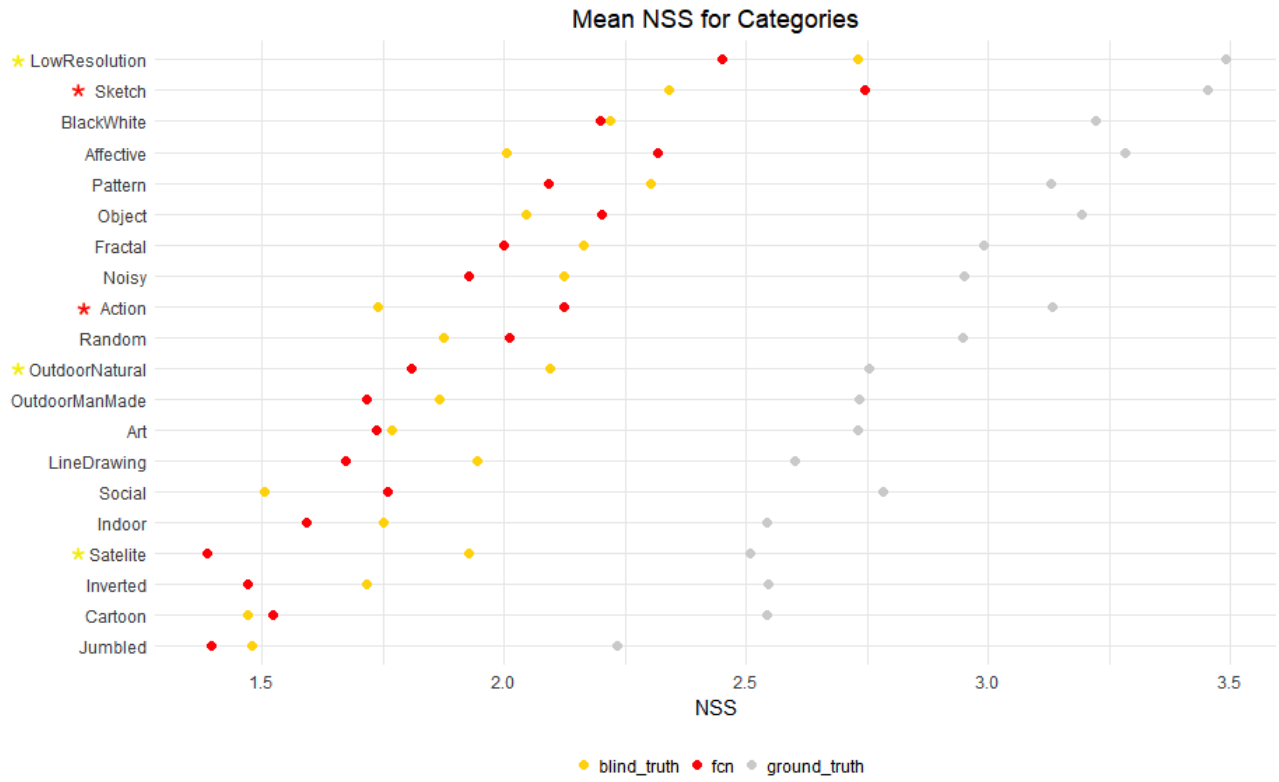


Figure 3: Mean NSS for each category of stimuli. The results were obtained using different methods (blind truth, fcn, ground truth). The significant differences between the Blind Truth and FCN models are marked with a star – yellow when the NSS for Blind Truth > FCN, red when FCN > Blind Truth

3.1. Conclusion

The main goal of the project has been accomplished. A deep neural network capable of predicting saliency maps was built. However, compared to the least effort solution (blind truth), the FCN network yield significantly better results only for 2 of 20 categories.

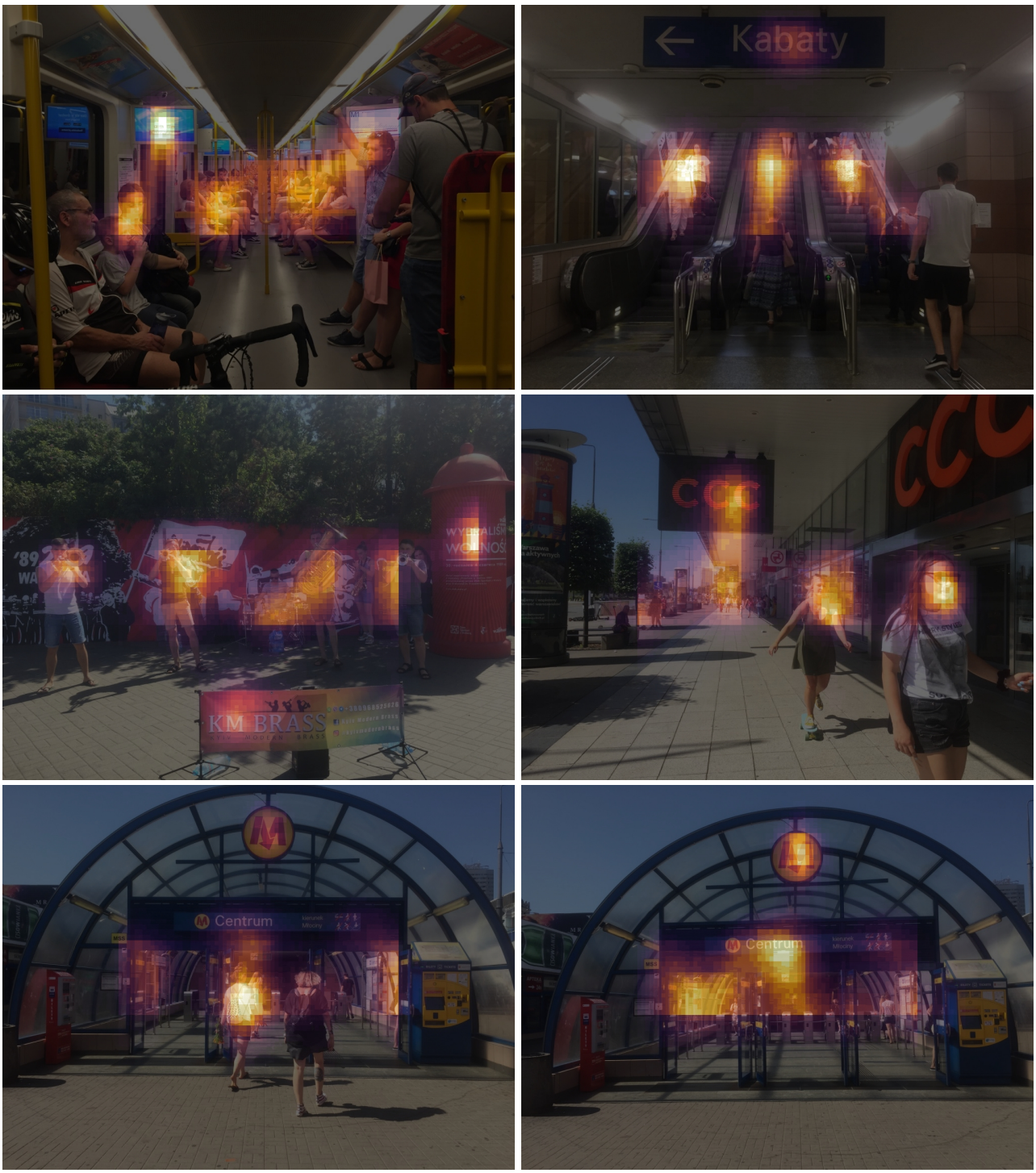


Figure 4. Saliency maps calculated by the FCN model for the photographs taken by the author.