From early biological models to CNNs: do they look where humans look?

Marinella Cadoni, Andrea Lagorio and Enrico Grosso
Computer Vision Laboratory
University of Sassari, Italy

Jia Huei, Tan and Chee Seng, Chan
Center of Image and Signal Processing
University of Malaya, Malaysia

Introduction

Early hierarchical computational visual models as well as recent deep neural networks have been inspired by the functioning of the primate visual cortex system. Considering the ability humans have to select high semantic level regions of a scene, the question whether neural networks can match this ability, and if similarity with humans attention is correlated with neural networks performance naturally arise.

To address this question, we propose a pipeline to select and compare sets of feature points that maximally activate individual networks units to human fixations. We extract features from a variety of neural networks, from early hierarchical models such as HMAX up to recent deep convolutional neural networks to compare them to human fixations. Experiments over the ETD database show that human fixations correlate with CNNs features from deep layers significantly better than with random sets of points, while they do not with features extracted from the first layers of CNNs, nor with the HMAX features, which seem to have low semantic level compared with the features that respond to the automatically learned filters of CNNs. It also turns out that there is a correlation between CNN's human similarity and classification performance.

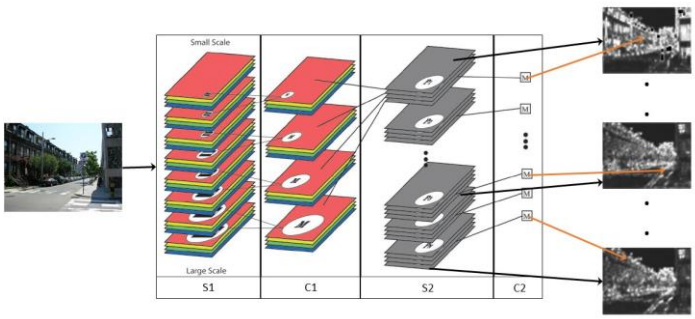
Human fixations and neural networks feature points.

From early biological models to CNNs: do they look where humans look?

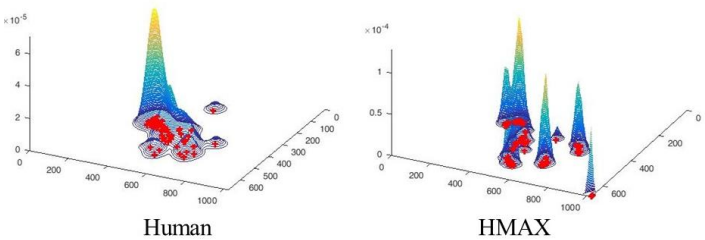
Marinella Cadoni, Andrea Lagorio and Enrico Grosso
Computer Vision Laboratory
University of Sassari, Italy

Jia Huei, Tan and Chee Seng, Chan
Center of Image and Signal Processing
University of Malaya, Malaysia

Scheme of the features extraction from HMAX



KDE estimation of the human fixations and the HMAX features



Interest area comparison

To evaluate the correlation between human fixations and interest points, we compare their density maps using three similarity indexes:

- **Bray-Curtis similarity (BC)**
- **Jensen-Shannon similarity (JS)**
- **Spearman rank correlation coefficient (ρ)**

Interest area comparison

Given an image **I** and two probability density functions $\mathbf{f_1}$ and $\mathbf{f_2}$

$$BC_{1,2} = 1 - \frac{\sum_{i=1}^{n} |f_1(x_i) - f_2(x_i)|}{\sum_{i=1}^{n} f_1(x_i) + f_2(x_i)}$$

$$JS_{1,2} = 1 - \sum_{x_i} (f_1(x_i) - f_2(x_i)) \, log \frac{f_1(x_i)}{f_2(x_i)}$$

Similarity indexes between all features and human fixations

Model	Layer	$BC_{f_{F_j} f_H}$	$JS_{f_{F_j} f_H}$	$\rho_{f_{F_j} f_H}$
		R=21.26%	R=34.11%	
HMAX		15.24%	24.81%	18.47
AlexNet	$C1$	25.51%	38.95%	31.36
	$C2$	31.61%	48.59%	29.66
	$C3$	34.45%	53.05%	37.19
	$C4$	34.47%	52.94%	36.38
	$C5$	34.09%	52.49%	35.39
VGG-19	$C1$	14.03%	22.99%	14.47
	$C2$	20.51%	32.12%	21.94
	$C3$	20.24%	47.91%	20.34
	$C4$	33.46%	50.05%	33.91
	$C5$	35.18%	53.44%	37.16
ResnetV2-50	$C1$	22.18%	**34.27%**	25.20
	$b1$	32.51%	49.42%	30.24
	$b2$	35.33%	54.24%	41.12
	$b3$	29.42%	47.09%	36.81
	$b4$	29.09%	46.71%	31.75
InceptionV3	$c1$	27.03%	41.27%	32.96
	$c2$	26.51%	40.40%	28.68
	$c3$	**21.87%**	**33.38%**	21.34
	$c4$	**21.87%**	**33.53%**	19.49
	$c5$	**21.88%**	**34.93%**	12.99
	$c6$	27.54%	44.63%	5.53
	$c7$	25.62%	42.35%	-6.14
Densenet-201	$C1$	**21.66%**	**33.53%**	24.25
	$C2$	29.82%	45.24%	27.28
	$C3$	34.65%	52.82%	32.24
	$C4$	32.36%	50.92%	30.16
	$C5$	29.02%	46.60%	32.20
EfficientNet-b7	$b1$	28.27%	42.95%	36.69
	$b2$	34.85%	52.64%	40.12
	$b3$	32.42%	50.17%	32.59
	$b4$	31.21%	49.01%	28.63
	$b5$	33.37%	51.42%	31.15
	$b6$	37.71%	56.98%	52.33
	$b7$	37.37%	56.36%	44.92

Similarity indexes between HMAX and DNNs features

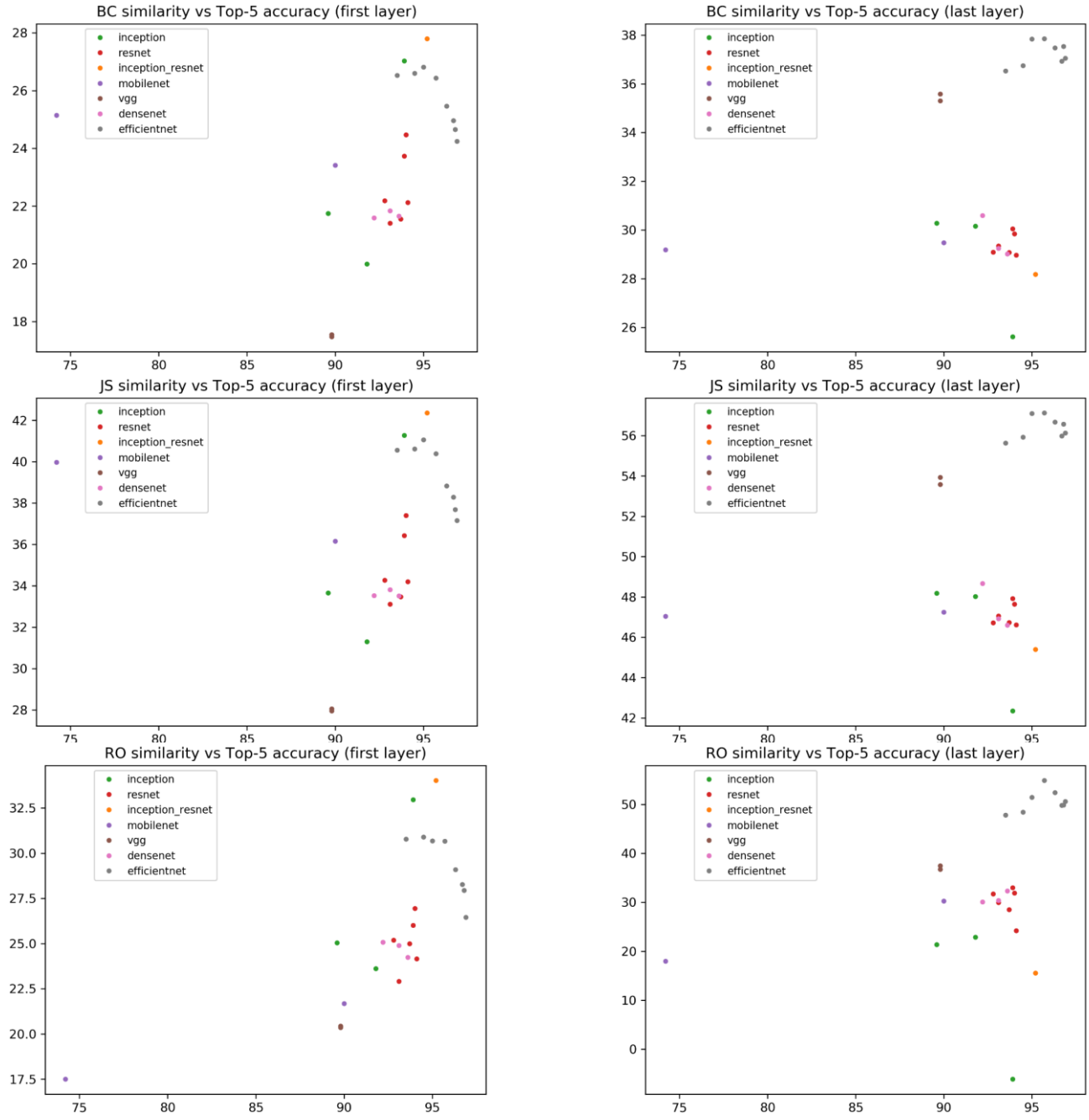
Model	Layer	$BC_{f_{F_j} f_{HMAX}}$	$JS_{f_{F_j} f_{HMAX}}$	$\rho_{f_{F_j} f_{HMAX}}$
		R=9.18%	R=17.12%	
AlexNet	$C1$	31.12%	46.84%	30.96
	$C2$	21.10%	36.37%	26.93
	$C3$	18.44%	33.31%	28.51
	$C4$	18.80%	33.63%	28.04
	$C5$	17.49%	31.74%	24.84
VGG-19	$C1$	24.16%	36.04%	18.85
	$C2$	34.65%	50.27%	28.53
	$C3$	32.06%	47.39%	27.88
	$C4$	25.41%	41.64%	31.23
	$C5$	17.35%	31.24%	24.38
ResnetV2-50	$C1$	29.12%	43.51%	27.17
	$b1$	25.43%	42.23%	31.86
	$b2$	18.53%	33.57%	32.60
	$b3$	12.59%	24.73%	19.69
	$b4$	12.37%	24.38%	16.49
InceptionV3	$c1$	23.30%	37.03%	25.61
	$c2$	30.15%	46.25%	31.41
	$c3$	28.42%	43.56%	29.29
	$c4$	29.09%	44.70%	29.16
	$c5$	19.23%	32.62%	20.88
	$c6$	14.16%	26.72%	13-27
	$c7$	12.05%	23.76%	7.28
Densenet-201	$C1$	27.75%	41.65%	25.93
	$C2$	30.29%	47.61%	31.54
	$C3$	22.38%	38.62%	30.40
	$C4$	15.05%	28.55%	23.56
	$C5$	12.46%	24.53%	18.78
EfficientNet-b7	$b1$	25.06%	39.63%	28.59
	$b2$	20.25%	34.98%	27.90
	$b3$	17.28%	31.02%	23.20
	$b4$	15.51%	28.73%	20.39
	$b5$	16.18%	29.37%	19.43
	$b6$	14.86%	27.76%	21.30
	$b7$	14.91%	27.66%	19.39

From early biological models to CNNs: do they look where humans look?

Marinella Cadoni, Andrea Lagorio and Enrico Grosso
Computer Vision Laboratory
University of Sassari, Italy

Jia Huei, Tan and Chee Seng, Chan
Center of Image and Signal Processing
University of Malaya, Malaysia

Scatter plots of human similarity (vertical axis) and performance (horizontal axis)



Correlation between human similarity and performance

	BC	JS	ρ
First Layer	0.63	0.61	0.76
Top Layer	0.38	0.37	0.54

Conclusions

Experimental results demonstrate that responses from the first filters do not correlate with human fixations, and HMAX features seem to be equivalent to the features extracted from the first layers of deep CNNs. Features from deep layers, on the contrary, correlate with human fixations. Similarity with humans is correlated with CNNs performance.