Supplementary Materials for: *Natural images are more informative for interpreting CNN activations than synthetic feature visualizations*

1 Details on Experimental Setup

The selection of feature maps used in Experiment I is shown in Table 1; the selection of feature maps used in Experiment II is shown in Table 2.

1.1 Selection of Feature Maps

Layer	Branch	Feature Map	Layer	Branch	Feature Map
mixed3a	1×1	25	mixed4d	1×1	95
	3×3	189		3×3	342
	5×5	197		5×5	451
	Pool	227		Pool	483
	Pool*	230		Pool*	516
mixed3b	1×1	64	mixed4e	1×1	231
	3×3	178		3×3	524
	5×5	390		5×5	656
	Pool	430		Pool	816
	Pool*	462		Pool*	809
mixed4a	1×1	68	mixed5a	1×1	229
	3×3	257		3×3	278
	5×5	427		5×5	636
	Pool	486		Pool	743
	Pool*	501		Pool*	720
mixed4b	1×1	45	mixed5b	1×1	119
	3×3	339		3×3	684
	5×5	438		5×5	844
	Pool	491		Pool	1007
	Pool*	465		Pool*	946
mixed4c	1×1	94			
	3×3	247			
	5×5	432			
	Pool	496			
	Pool*	449			

Table 1: Feature maps analyzed in Experiment I. For each of the 9 layers with an Inception module, one randomly chosen feature map per branch (1×1, 3×3, 5×5 and Pool) and one additional hand-picked feature map (highlighted with *) are used.

Layer	Branch	Feature Map for Batch Block (A-D)			
		A	B	C	D
mixed3a	1×1	25	14	12	53
	3×3	189	97	171	106
	5×5	197	203	212	204
	Pool	227	238	232	247
mixed4a	1×1	68	33	45	17
	3×3	257	355	321	200
	5×5	427	425	429	423
	Pool	486	497	478	506
mixed4c	1×1	94	53	59	95
	3×3	247	237	357	209
	5×5	432	402	400	416
	Pool	496	498	473	497
mixed4e	1×1	231	83	6	89
	3×3	524	323	401	373
	5×5	656	624	642	620
	Pool	816	755	724	783
mixed5b	1×1	119	14	266	300
	3×3	684	592	657	481
	5×5	844	829	839	875
	Pool	1007	913	927	903

Table 2: Feature maps analyzed in Experiment II. Four sets of feature maps (batch blocks A to D) are sampled: For every second layer with an Inception module (5 layers in total), one feature map is randomly selected per branch of the Inception module (1×1, 3×3, 5×5 and Pool). For the practice, catch and intuitiveness trials additional randomly chosen feature maps are used.

Subject	**Order of presentation schemes (0-3) and batch-blocks (A-D)**	**Batches**		**Order of synthetic and natural**
		Practice	**Main**	
1	0 (A) 1 (B) 2 (C) 3 (D)			
2	0 (B) 2 (D) 1 (C) 3 (A)		natural: 1	natural - synthetic
3	3 (B) 1 (D) 2 (A) 0 (C)		synthetic: 2	
4	3 (C) 2 (B) 1 (A) 0 (D)			
5				
6	see subject 1-4		natural: 3	synthetic - natural
7		0	synthetic: 4	
8				
9				
10	see subject 1-4		natural: 5	natural - synthetic
11			synthetic: 6	
12				
13	see subject 1-4		natural: 7 synthetic: 8	synthetic - natural

Table 3: **Counter-balancing of conditions in Experiment II.** In total, 13 naive and 10 lay participants are tested. Each "batch blocks" contains 20 feature maps (sampled from five layers and all Inception module branches). Batches indicate which batch number the natural query (and reference images) are taken from.

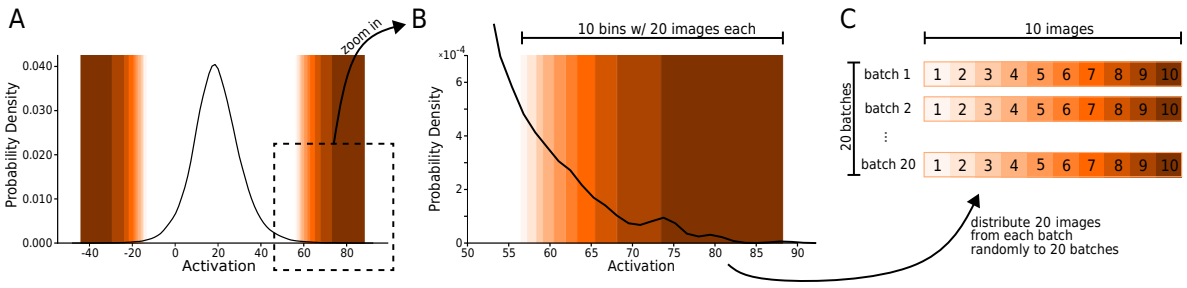

Figure 1: Sampling of natural images for Experiment I and II. **A:** Distribution of activations. For an example channel (mixed3a, kernel size 1×1, feature map 25), the smoothed distribution of activations for all $50,000$ ImageNet validation images is plotted. The natural stimuli for the experiment are taken from the tails of the distribution (shaded background). **B:** Zoomed-in tail of activations distribution. In the presentation schemes with 9 images, 10 bins with 20 images each are created (10 because of 9 reference plus 1 query image). **C:** In order to obtain 20 batches with 10 images each, the 20 images from one bin are randomly distributed to the 20 batches. This guarantees that each batch contains a fair selection of extremely activating images. The query images are *always* sampled from the most extreme bins in order to give the best signal possible. In the case of the presentation schemes with 1 reference image, the number of bins in B is reduced to 2 and the number of images per batch in C is also reduced to 2.

2 Screenshots

(a) Screen at the beginning of a trial. The question is which of the two natural images at the center of the screen also strongly activates the CNN feature map given the reference images on the side.

(b) Screen including a subject's answer visualized by black boxes around the image and the confidence level. A subject indicates which natural image in the center would also be a strongly activating image by clicking on the number corresponding to his/her confidence level (1: not confident, 2: somewhat confident, 3: confident). The time until a participant selects an answer is recorded ("reaction time").

(c) Screen including a subject's answer (black boxes) and feedback on which image is indeed also a strongly activating image (green box).

Figure 2: Screenshots of the experimental setup (Forward Simulation Task). The progress bar at the bottom of the screen indicates the progress within one block of trials.

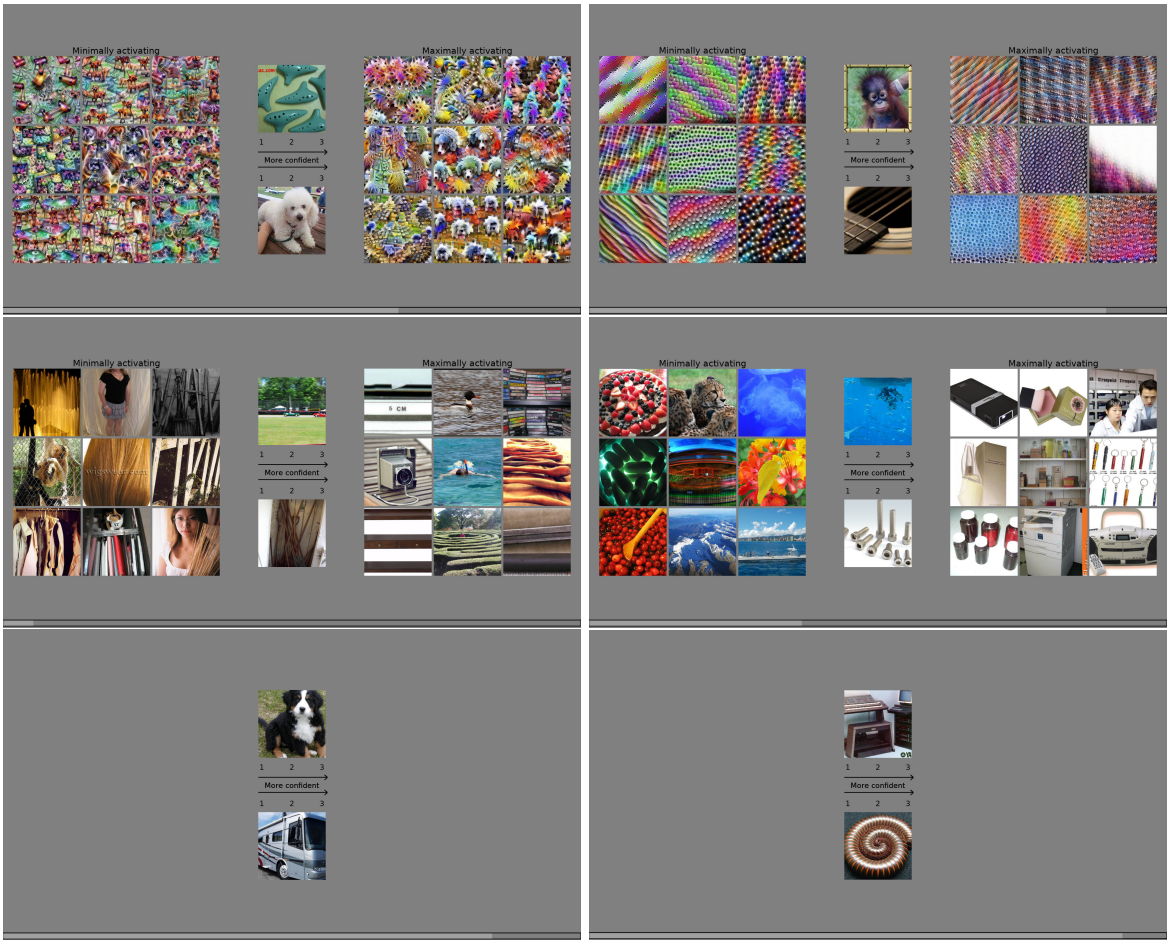

Figure 3: Experiment I: Example trials of the three reference images conditions: synthetic reference images (first row), natural reference images (second row) or no reference images (third row). The query images in the center are always natural images.

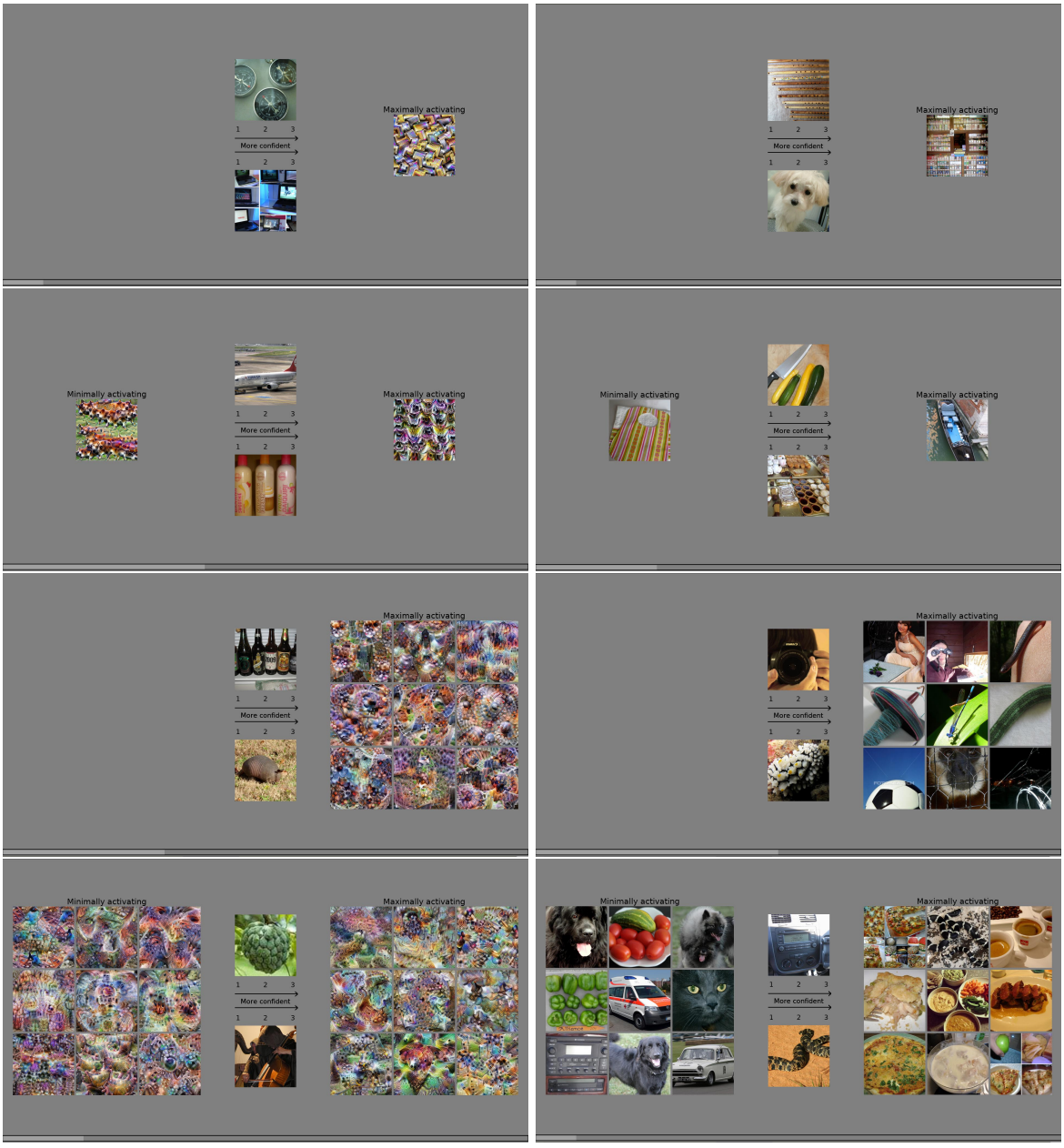

Figure 4: Experiment II: Example trials of the four presentation conditions: Max 1, Min+max 1, Max 9, Min+Max 9. The left column contains the synthetic reference images, the right column contains the natural reference images.

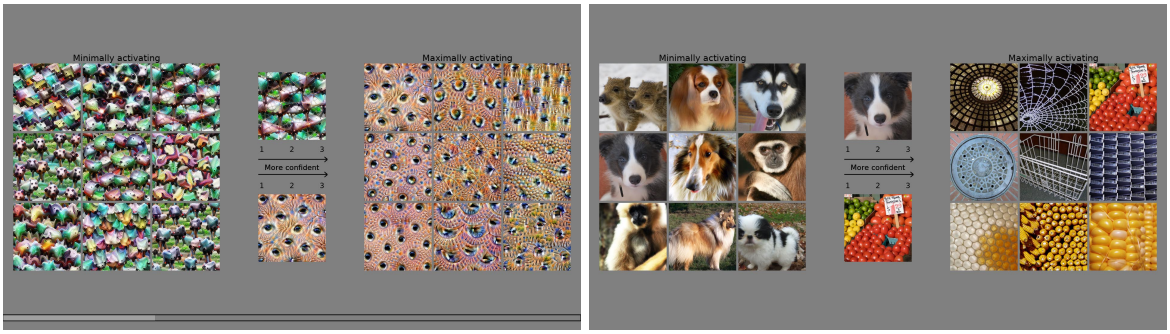

Figure 5: Catch trials. An image from the reference images is copied as a query image, which makes the answer obvious. The purpose of these trials is to integrate a mechanism into the experiment which allows to check post-hoc whether a participant was still paying attention.

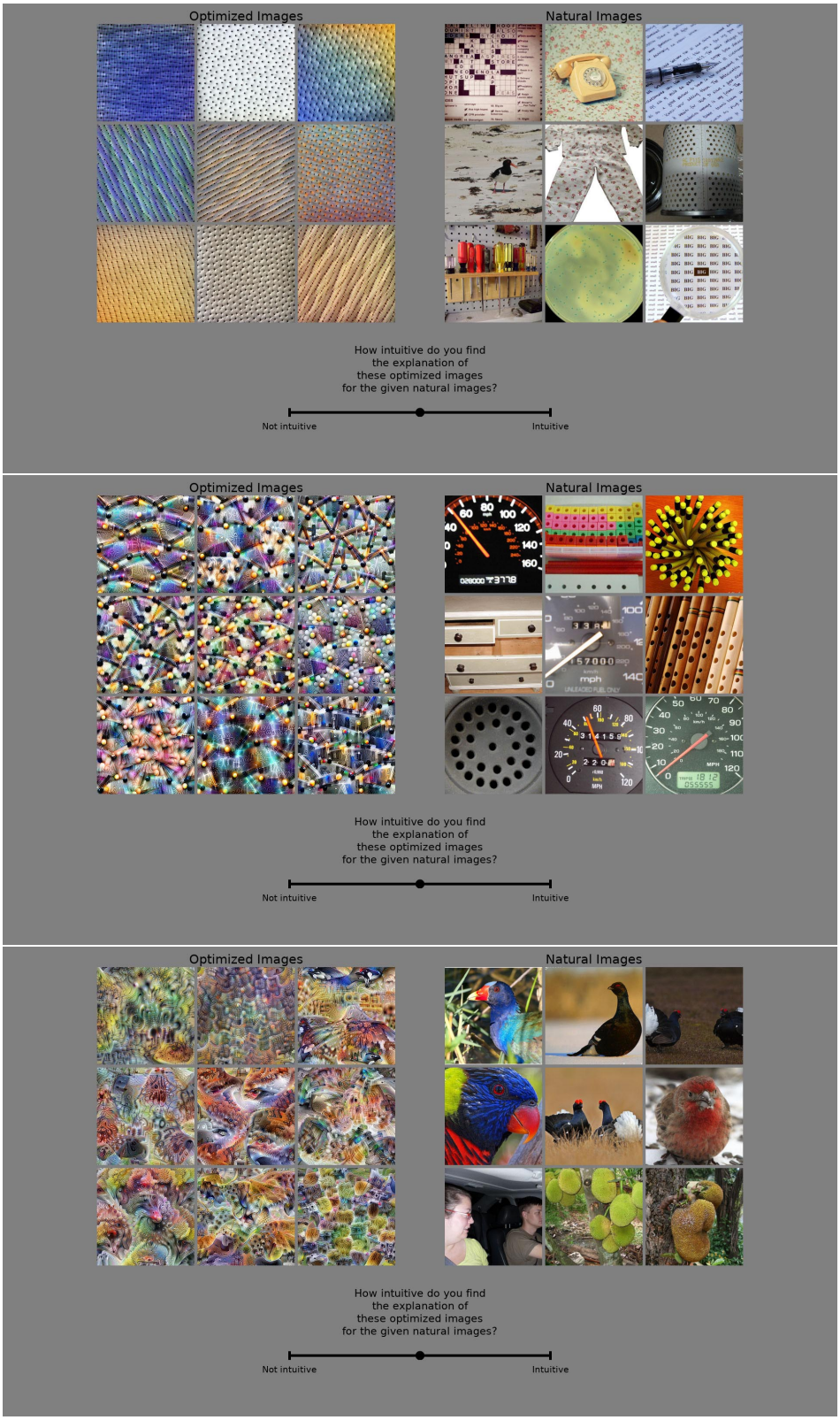

Figure 6: Trials for intuitiveness judgment. The tested feature maps are layer mixed3a (channel 43), mixed4b (channel 504) and mixed 5b (channel 17). They are the same in Experiment I and in Experiment II.