eNeuro

Research Article: New Research* | *Cognition and Behavior

Differential involvement of EEG oscillatory components in sameness vs. spatial-relation visual reasoning tasks

https://doi.org/10.1523/ENEURO.0267-20.2020

Cite as: eNeuro 2020; 10.1523/ENEURO.0267-20.2020

Received: 15 June 2020
Revised: 20 October 2020
Accepted: 21 October 2020

Differential involvement of EEG oscillatory components in sameness vs. spatial-relation visual reasoning tasks

Andrea Alamia[1], Canhuang Luo[1], Matthew Ricci[2], Junkyung Kim[2], Thomas Serre[2,3] and Rufin VanRullen[1,3]

rufin.vanrullen@cnrs.fr

[1]CerCo, CNRS Université de Toulouse, Toulouse 31055 (France)
[2] Department of Cognitive, Linguistic & Psychological Sciences; Carney Institute for Brain Science; Brown University, Providence, RI 02912, USA.
[3]ANITI, Université de Toulouse, Toulouse 31055 (France)

Abstract

The development of deep convolutional neural networks (CNNs) has recently led to great successes in computer vision and CNNs have become de facto computational models of vision. However, a growing body of work suggests that they exhibit critical limitations beyond image categorization. Here, we study one such fundamental limitation, for judging whether two simultaneously presented items are the same or different (SD) compared to a baseline assessment of their spatial relationship (SR). In both human subjects and artificial neural networks, we test the prediction that SD tasks recruit additional cortical mechanisms which underlie critical aspects of visual cognition that are not explained by current computational models. We thus recorded EEG signals from human participants engaged in the same tasks as the computational models. Importantly, in humans the two tasks were matched in terms of difficulty by an adaptive psychometric procedure: yet, on top of a modulation of evoked potentials, our results revealed higher activity in the low beta (16-24Hz) band in the SD compared to the SR conditions. We surmise that these oscillations reflect the crucial involvement of additional mechanisms, such as working memory and attention, which are missing in current feed-forward CNNs.

Keywords: Visual reasoning, spatial relationship, EEG oscillations, ERPs, deep neural networks.

Significance statement

Convolutional neural networks (CNNs) are currently the best computational models of primate vision. Here, we independently confirm prior results suggesting that CNNs can learn to solve visual reasoning problems involving spatial relations much more easily than problems involving sameness judgments. We hypothesize that these results reflect different computational demands between the two tasks and conducted a human EEG experiment to test this hypothesis. Our results suggest a significant difference – both in evoked potentials and in the oscillatory dynamics– of the EEG signals measured from human participants performing these two tasks. We interpret this difference as the signature for the fundamental involvement of recurrent mechanisms implementing cognitive functions such as working memory and attention.

1. Introduction

The field of artificial vision witnessed an impressive boost in the last few years, driven by the striking results of deep convolutional neural networks (CNNs). Such hierarchical neural networks process information sequentially – through a feedforward cascade of filtering, rectification and normalization operations. The accuracy of these architectures is now approaching – sometimes exceeding – that of human observers on key visual recognition tasks including object (He, Zhang, Ren, & Sun, 2016) and face recognition (P. J. Phillips et al., 2018). These advances suggest that purely feedforward mechanisms suffice to accomplish remarkable results in object categorization, in line with previous experimental studies on humans (VanRullen & Thorpe, 2001) and animals (Hollard & Delius, 1982; Vogels, 1999). However, despite the remarkable accuracy reached in these recognition tasks, the limitations of CNNs are becoming increasingly evident (see Serre, 2019 for a recent review). Beyond image categorization tasks, CNNs appear to struggle to learn to solve relatively simple visual reasoning tasks otherwise trivial for the human brain (Kim, Ricci, & Serre, 2018; Stabinger, Rodríguez-Sánchez, & Piater, 2016). A recent study (Kim et al., 2018) thoroughly investigated the ability of CNN architectures to learn to solve various visual reasoning tasks, and found an apparent dichotomy between two sorts of problems: on the one hand, tasks that require judging the spatial relations between items (Spatial Relationship – SR); on the other, those that require comparing items (Same-Different – SD). Importantly, Kim and colleagues demonstrated that CNNs can more easily learn the first class of problems compared to the second one.

This prompts the question of how biological visual systems handle such tasks so efficiently. Kim et al. (2018) suggest that SR and SD tasks tap into distinct computational mechanisms, thus leading to the prediction that different cortical processes are also involved when humans perform the two tasks: SR tasks can be successfully solved by feedforward processes, whereas SD tasks seem to require additional computations, such as working memory and attention. Here, we tested this hypothesis in two steps: first, we confirmed and extended Kim's results by comparing the performance of CNNs on an experiment in which we directly contrasted SD and SR tasks on the same stimulus set. Second, we recorded electrophysiological responses (EEG) in healthy human participants for the same experiment, after having matched the difficulty level via an adaptive psychometric procedure. We hypothesized that the additional computations required by the SD task, as compared to SR tasks, would elicit differences in evoked potentials (e.g. P300 modulations, which have been related to attentional mechanisms (Nash & Fernandez, 1996)) and brain rhythms related to working memory (such as beta-band oscillations (Benchenane, Tiesinga, & Battaglia, 2011; Lundqvist, Herman, Warden, Brincat, & Miller, 2018)). We found indeed that, in addition to a variation in evoked potentials, the SD task elicited higher activity in specific beta-band oscillatory components in the occipital-parietal areas, which are typically associated with attention- and memory-related processes. We emphasize that the goal of the present study was not to identify the precise neural computations involved in the two tasks (which would naturally require a broader experimental set-up than a single EEG study), but rather to validate the hypothesis that SD involves additional computations relative to SR (even when the two tasks are equally difficult). We hope that this demonstration can be a first step towards characterizing the processes taking place in visual cortex during visual reasoning tasks, and designing more reliable and more human-like computational models.

2. Materials & Methods

2.1 Participants and pilot experiment

Twenty-eight participants (aged 21–34 years old with a mean age of 26.6 ± 3.7, 11 women, 5 left-handed), volunteered to join the experiment. All subjects reported normal or corrected to normal vision and had no history of epileptic seizures or neurological disorders. Participants were pooled in two groups of 14 each: one group performed a pilot experiment, while the second one was tested on a final version of the task. The only difference between the pilot and the main study was the QUEST adaptive procedure used to match the difficulty level between conditions, which was not implemented in the pilot experiment. However, in both studies we found the very same result (see below, specifically fig. 4 and 5). In the main experiment, we kept the same number of participants to replicate the effect, after having removed the behavioral difference in task difficulty via the QUEST algorithm. This study complies with the guidelines of the research center where it was carried out, and the protocol was approved by an external committee (ethics approval number N° 2016-A01937-44). All participants gave written informed consent before starting the experiment, in accordance with the Declaration of Helsinki, and received monetary compensation for their participation.

2.2 Experimental design

The experiment was composed of 16 experimental blocks of 70 trials each, with a total duration of about 1 hour. Each trial lasted ~2 seconds (Fig. 1A): 350ms after the onset of a black fixation cross (0.6° width), 2 shapes were displayed for 30ms on opposite sides of the screen, distant 2*ρ from each other with an angle of $\pm(45° + \theta)$ with respect to the horizontal midline (ρ being the distance from the center of the screen, and θ the angular difference with the diagonal; see Fig. 1B). Each shape was selected from a subset of 36 hexominoes, a geometric figure composed of 6 contiguous squares (see Fig. 1B) One second after the onset of the hexominoes, the fixation cross turned blue, cuing participants to respond. In half of the blocks, participants had to report whether the two shapes were the same or different (Same-Different – SD condition); in the remaining blocks participants had to judge whether the two stimuli were aligned more horizontally or vertically (Spatial Relation – SR condition). Shapes were displayed at opposite sides of the screen along two main possible orientation axes sampled at random for every trial (either 45° and 225° or -45° and -225°). Both stimuli positions were jittered by a random offset Δx and Δy in both the x and y axis and a rotation θ from the main axis. The same offsets were applied to both shapes, so they did not affect the angle between stimuli. The aim of such offsets was to prevent participants in the SR condition from determining the configuration of the two stimuli (orientation task) by merely judging the position of a single stimulus: without the random offsets, considering for example the top-right corner position, if the item were below/above the (imaginary) screen diagonal line, the overall orientation would be horizontal/vertical, without the need to consider the position of the corresponding bottom-left item. The offset then compelled participants to consider the relative position of both hexominoes at once. Importantly, in the main experiment (compared to the pilot experiment) the difficulty of the two tasks was controlled by an adaptive psychometric procedure (QUEST method, Watson & Pelli, 1983), which varied the eccentricity of the two stimuli ρ (in the SD blocks) or θ (in the SR blocks) to maintain an overall accuracy level of 80% throughout the whole experiment. In fact, larger (smaller) values of ρ made the stimuli more (less) eccentric and the task more (less) difficult; similarly, smaller (larger) values of θ set the stimuli closer to (farther from) the 45° diagonal line, making the task more (less) difficult. We modified one parameter per condition (i.e., per block), while the other was kept constant (using the same value as in the preceding block). After participants responded, they received feedback on their performance: the fixation cross turned green (red) in case of a correct (incorrect) answer. Throughout the experiment the condition blocks were alternated, the first block being the SD condition for all participants. Before starting the first block, participants performed one training block per condition. The purpose of this training was 1) to familiarize participants with the experimental conditions, 2) to initialize the ρ and θ parameters in the QUEST method for the first experimental block (initial values were respectively ρ = 5.4° of visual angle and θ=6° of rotation). All experiments were performed on a cathode ray monitor, positioned 57 cm from the subject, with a refresh rate of 160 Hz and a resolution of 1280 × 1024 pixels. The experiment was coded in MATLAB using the Psychophysics Toolbox (Brainard, 1997). The stimuli were presented in black on a gray background. Throughout the experiment we recorded EEG signals.

2.3 EEG recording and pre-processing

We recorded brain activity using a 64-channel active BioSemi electro-encephalography (EEG) system (1,024 Hz digitizing rate, 3 additional ocular electrodes). The pre-processing was performed in MATLAB using the EEGlab toolbox (Delorme & Makeig, 2004). First, the data was downsampled to 256 Hz. A notch filter [47Hz - 53Hz] was then applied to remove power line artifacts. We applied an average-referencing and removed slow drifts by applying a high-pass filter (>1 Hz). We created the data epochs aligning the data to the onset of the fixation cross. Finally, we performed an ICA decomposition in order to remove components related to eye movements and blink artifacts: we visually inspected the data and removed from 2 to 5 components per subject with a conservative approach (we removed only components in the frontal regions clearly related to eye movements' activity).

2.4 Computational modeling and code accessability

We extended a previous computational study (Kim et al., 2018) from which we chose the parameters of the convolutional feedforward network trained on the SD and SR tasks. Each task was run 10 times, randomly initializing the networks' parameters and the stimuli used in the training and test set. The network was fed with 50x80 pixel images. Two hexominoes (width and height of 2 to 5 pixels) were placed at opposite sides of the screen (see Fig. 1A and 'Experimental design'). The dictionary of hexominoes was composed of 35 items, which were randomly split between a training (30 items) and a test set (5 items) at each iteration. Both the training, validation and test sets were composed of 1,000 stimuli (i.e. different combinations of the hexominoes, with slightly different eccentricity and/or offset relative to the diagonal). The network consisted of 6 convolutional layers. Each layer contained 4 channels of size 2x2, with stride of 1. All convolutional layers used a ReLu activation function with stride of 1 and were followed by pooling layers with 2x2 kernels and a stride of 1. Eventually, two fully connected layers with 128 units preceded a two-dimensional classification layer with a sigmoid activation function. As a regularizer we set a dropout rate of 0.3 in each layer of the network. We used binary cross-entropy as a loss function, the Adaptive Moment Estimation (Adam) optimizer (Kingma & Ba, 2015) and a learning rate of 10e-4. Each simulation was run over 70 epochs with batch size of 50. All simulations were run in TensorFlow (GoogleResearch, 2015). The Siamese network had the same exact convolutional architecture as described above; additionally, the difference between features-vectors of each separate item (computed on an input image where this item was shown alone) was fed to the classifier to perform the SD task. All networks count ~7e06 parameters. All the code and data required to replicate the simulations are available at a github repository (https://github.com/artipago/SD-SR). The code has been run on a Window PC on Python using the "Tensorflow", "Keras", "Scipy" and "Numpy" libraries.

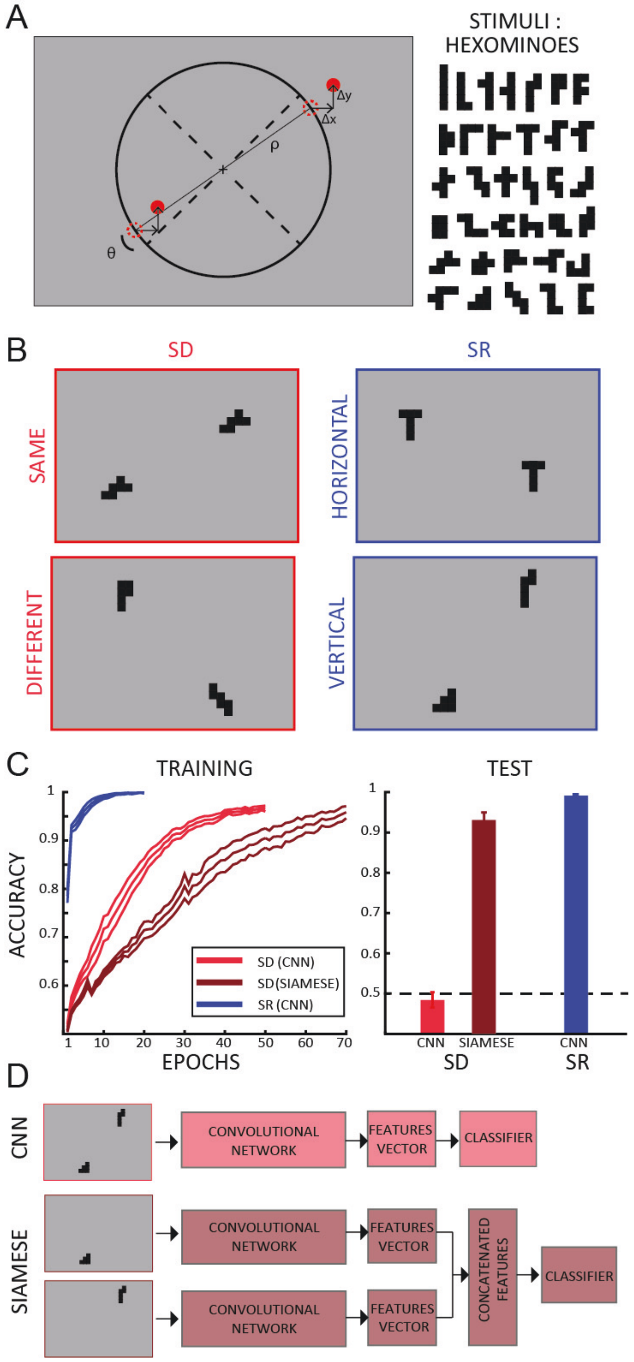

Fig.1: **Stimuli and simulation results**. A) The stimuli were the same in the simulations and in the human experiments. The items were displayed at opposite sides of the screen (either 45° and 225° or -45° and -225°). Both item positions were jittered by a random amount in both the x and y axis (Δx and Δy in the picture) to make the task non-trivial for human participants (i.e. preventing participants from performing the SR task considering only the position of one item, thus ignoring the spatial relationship between the two items). The items used are hexominoes (right panel). Minimum and maximum item height and width are 1.2° – 3.6° and

1.2° – 2.7° of visual angle respectively, and 2 to 5 pixels used for the simulations (image size was 50 x 80 pixels).B) Example of stimuli position for the same-different task (SD - left column) and spatial relation task (SR - right column). For the sake of illustration the ratio between the screen and hexominoes size has been modified (stimuli here look bigger than in the real experiment). C-D) Accuracy of the CNN network on the same-different (SD; light red) and spatial relationship (SR; blue) tasks, and of a Siamese network trained on the SD task (dark red). The Siamese network mimics segmentation in a feedforward network, by separating the items in two distinct channels of the network (see panel D). The left panel shows the training curves for each network (accuracy over epochs during training); we stopped the training when the validation accuracy reached 90%. In the right panel we show the training accuracy at the last epoch and the test accuracy. The latter was evaluated using novel items never used for training, and it reveals that the CNN seems to only learn the required rule for the SR but not for the SD task, as shown in a previous study. Conversely, the Siamese network (CNN with segmentation) can solve the SD task, demonstrating that segmentation can allow the CNN to successfully accomplish this task. In both panels we show average values ± SE over 10 repetitions using different random initializations.

2.5 Statistical analysis – behavior

We analyzed both accuracy and reaction times (RT) by means of Bayesian ANOVA, considering the block condition (SR and SD, see above) as independent variables and the trial condition (whether the stimuli were same or different, or more horizontally or vertically aligned). The result of such analysis provides a Bayes Factors (BF), which quantifies the ratio between statistical models given the data. Throughout the paper, all BFs reported correspond to the probability of the alternative hypothesis over the null hypothesis (indicated as BF_{10}). Practically, a large BF (~BF>5) provides evidence in favor of the alternative hypothesis (the larger the BF the stronger the evidence), whereas low BF (~BF<0.5) suggests a lack of effect (Masson, 2011; Smith, 2001). We performed all Bayesian analyses in JASP (JASP Team, 2018; Love et al., 2015).

2.6 Statistical analysis – electrophysiology

Regarding the EEG recording we performed 2 analyses: one in the time domain measuring Evoked Related Potentials – ERPs, and the other one in the frequency domain using a time-frequency transform. In the first case, we considered the ERPs recorded from 7 midline electrodes (i.e., Oz, POz, Pz, CPz, Cz, FCz and Fz). After subtracting the baseline activity recorded during the 350ms before stimuli onset, we averaged the signals from the SD and SR blocks respectively (i.e., 8 blocks for each condition). Finally, we tested whether the difference between these signals differed from 0 by means of a point-by-point 2-tailed t test with a false discovery rate (FDR) correction for multiple comparisons (Hochberg, 1995). Regarding the time-frequency analysis, we computed the power spectra by means of a wavelet transform (1–50 Hz in log-space frequency steps with 1-20 cycles). After baseline correction (i.e., dividing by the averaged activity of the 350ms prior to the onset of the fixation cross), for each participant, we computed the difference in decibel of the two conditions point by point, averaging over all electrodes. As in the ERP analysis, we performed a point-by-point 2-tailed t test to identify the time-frequency regions which were significantly different. We applied a cluster-based permutation to correct for multiple comparisons (Maris & Oostenveld, 2007). First, we identified clusters composed of t values t>3.5 (p<0.01), and for each one we computed the respective global sum. In order to estimate the null distribution over the combined t values, we performed the same procedure 500 times after shuffling the subject by subject SD-SR assignment. Eventually, we obtained the p values for each non-shuffled cluster given the null distribution. All EEG analyses were performed in Matlab; the wavelet transform was performed using the EEGlab toolbox (Delorme & Makeig, 2004).

3. Results

3.1 Computational modeling

We first extended the results by Kim et al. (2018) for our novel stimulus set: we trained two separate Convolutional Neural Networks (CNN) architectures to solve an SD and an SR task using a single stimulus set (Methods). The input to these networks was an image (50x80 pixels) in which two hexominoes (width and height of 2 to 5 pixels) were displayed at opposite sides of the screen (see Fig. 1A). The networks were trained to classify whether the two hexominoes were the same or not (SD task) or whether they were aligned more vertically or more horizontally with respect to the midline (SR task).

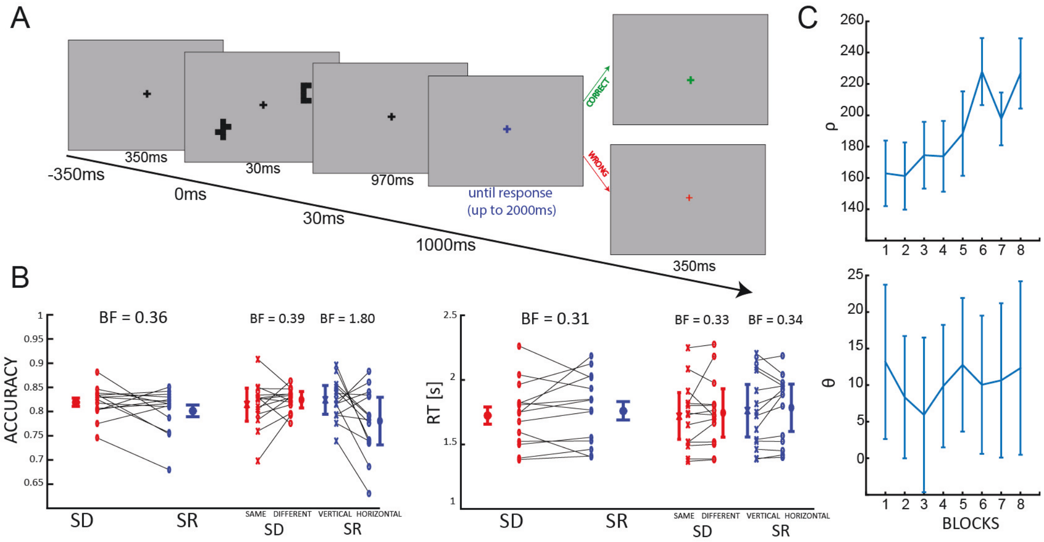

Fig.2: **Experimental design and human behavioral results**. A) At the beginning of each trial a black fixation cross was displayed for 350ms. After 2 stimuli were shown for 30ms, participants waited an additional 970ms before providing the answer. The response was cued by the fixation cross turning blue. After the response, the color of the fixation cross provided feedback: green if the response was correct, red otherwise. B) Humans performed the SD and SR tasks with comparable levels of performance. In the left and right panels are shown

the averages ± SE for accuracy and reaction times, respectively. Each pair of connected markers represent an individual subject. The results for the same-different (in red) and spatial relationship (in blue) conditions are further broken down for each condition separately (same-different and vertical-horizontal). BF indicates the Bayes factor against the null hypothesis (difference between the two conditions). C) Changes over blocks of ρ (the distance between the stimuli - left panel) and θ (the angle between the stimuli and the meridian - right panel) as adjusted by the QUEST algorithm.

We trained and tested the network on different sets of items (a training and test set, respectively) to assess the networks' ability to generalize beyond training data. We trained and tested the networks 10 times – randomly initializing networks parameters and training – test set split each time. We report the mean accuracy and standard deviation over these 10 repetitions in Fig. 1B. Our results are consistent with those from Kim et al. (2018): a CNN appears to be able to learn the abstract rule (as measured by the network's ability to generalize beyond the shapes used for training) for SR tasks much more easily than SD tasks. The effortless ability of humans and other animals (Daniel, Wright, & Katz, 2015; Wasserman, Castro, & Freeman, 2012) to learn SD tasks suggest the possible involvement of additional computations that are lacking in CNNs, possibly achieving items identification or segmentation (e.g. via attention and working memory). In order to verify that segmentation could be a missing ingredient for the SD task, we implemented a variant of the CNN with built-in segmentation properties, and tested it on the SD task (it is not necessary to test it on the SR task, because generalization performance is already at ceiling). The new network used a Siamese architecture (Bromley et al., 1994) in which each item is processed separately and eventually combined before being passed to a classifier. Therefore this model mimics the effect of selective attention and item segregation by feeding to the network each item separately. The Siamese network could achieve the same training performance on the SD task as the standard CNN (even though the training took more epochs), however the network was able to generalize to the test set, while the standard CNN test accuracy was at chance. This supports the idea that item segmentation or individuation abilities are needed to achieve the SD task. Next, we test the prediction that SD tasks in humans also require additional computational mechanisms than SR tasks, by recording EEG signals from a pool of 28 participants (14 of which were tested on a pilot experiment –fig. 5) performing the same SD and SR tasks.

3.2 Human behavior

A first pilot group of 14 participants performed the SD and SR tasks as described in Figure 2A, but without any procedure for adjusting task difficulty (i.e. the QUEST method). The same EEG oscillatory differences between the two tasks as in the main experiment were observed (fig 5); however, concomitant differences in behavioral task performance left open the possibility that the oscillatory effects were caused by differences in task difficulty (fig. 5A).

Therefore, we replicated the experiment on another group of 14 subjects, this time with an adaptive procedure to equate behavioral performance between the SD and SR tasks.

Participants (N=14) in this main experimental task completed 16 blocks using the same stimuli as those used to train CNNs (Fig.1): in half of the blocks they were asked to report whether the two hexominoes were the same or not (SD conditions), in the other half whether the hexominoes were more vertically or horizontally aligned (SR conditions). The two conditions were interleaved in a block design. Participants were required to answer after one second from stimulus onset in order to disentangle motor from visual components in the EEG recordings (Fig. 2A). The QUEST algorithm was used to assure that participants' accuracy was matched between the two tasks and remained constant throughout the whole experiment. This was done by adjusting two experimental parameters trial by trial (i.e., the hexominoes eccentricity in SD blocks, ρ, and the angle from the diagonal in SR blocks, θ; see Fig. 1A and 2C). Maintaining a comparable accuracy between the two tasks reduces the potential for confounds in the electrophysiological analysis due to differences in performance, vigilance or motivation. We confirmed the absence of any substantial behavioral difference between the SD and SR tasks (Fig. 2B) with a Bayesian ANOVA on both accuracy (BF_{10} = 0.361, error < 0.001%) and RT (BF_{10} = 0.317, error < 0.89%). In addition, we also investigated each condition separately (Fig. 2B), comparing the difference between 'same' and 'different' trials (in SD blocks) and 'vertical' and 'horizontal' trials (in SR blocks) in both RT and accuracy. All comparisons revealed overall no differences between tasks, except for the accuracy of vertical and horizontal trials in the SR condition, in which the BF proved inconclusive (accuracy: SD - BF_{10} = 0.39, error < 0.012%; SR - BF_{10} = 1.80, error < 0.001%; RT: SD - BF_{10} = 0.333, error < 0.01%; SR - BF_{10} = 0.34, error < 0.01%).

3.3 Human electrophysiology: evoked potentials

After having confirmed that performance was equal in the two tasks, we characterized the evoked potentials (EP) in each task. First, we estimated the difference between SR and SD conditions considering 7 midline electrodes (Fig.3). The results of a point-by-point t-test corrected for multiple comparisons revealed a significant difference in central and posterior electrodes (mostly Pz and CPz) between 250ms after the onset of the stimuli and the response cue, and the opposite effect in frontal electrodes (FCz and Fz) from 750ms to 1000ms, as confirmed by the topography (Fig.3). Overall these results indicate larger potentials in visual areas during the SD task than in the SR. Previous studies have shown a relation between EP amplitude (particularly P300 and late components) with attention (Itthipuripat, Cha, Byers, & Serences, 2017; Itthipuripat, Cha, Deering, Salazar, & Serences, 2018; Krusemark, Kiehl, & Newman, 2016; Van Voorhis & Hillyard, 1977) and visual working memory (Fabiani, Karis, & Donchin, 1986; Kok, 2001; McEvoy, 1998). Our results are thus consistent with a larger involvement of executive functions in the SD vs. SR task. In the following, we investigated whether this hypothesis is corroborated by corresponding oscillatory effects in the time-frequency domain in the main experiment.

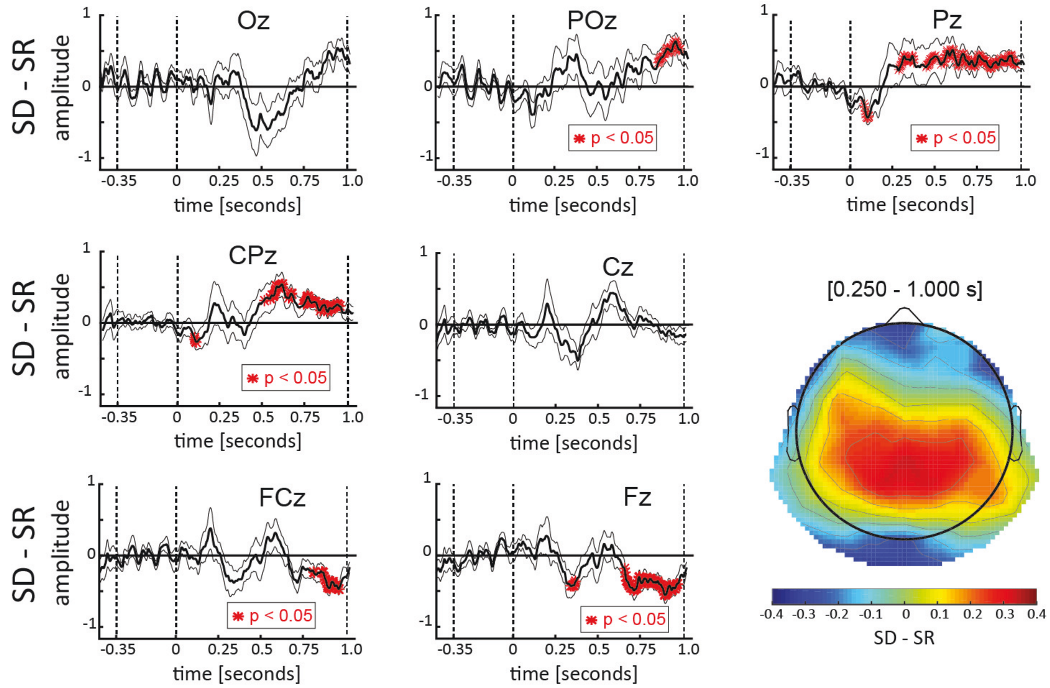

***Fig.3:* ERPs results**. Each panel represents the difference between ERPs elicited in the SD and SR conditions for the 7 midline electrodes (average ± SE). Shown in red are the points for which a significant difference was found against zero. The results reveal a significant difference from 250ms after stimuli onset until the response cue (at 1000ms) in central parietal regions, and an opposite effect after 750ms in frontal regions. In the bottom-right panel the topography, computed over the 250ms – 1000ms interval, confirmed a larger activity in the SD than in the SR condition (positive difference, warmer colors) in the central-parietal regions, and an opposite effect (negative difference, colder colors) in the frontal regions (which –although not significantly– also included occipital regions).

3.4 Human electrophysiology: time-frequency analysis

We performed a time-frequency analysis to try to identify differences between conditions observed in specific frequency bands commonly related to executive functions (e.g., visual working memory). For this purpose, we computed a baseline-corrected log-scale ratio between the two conditions (as shown in Fig. 4A), averaging over all electrodes. Remarkably, a point-by-point 2-tailed t-test corrected with cluster-based permutation test (Maris & Oostenveld, 2007) revealed a significantly larger activity in the low beta-band (16-24Hz) in the SD condition between 250 and 950ms after stimuli onset (Fig. 4B). We further quantify the magnitude of the effect by computing the effect size of a one sample t-test against zero averaging per each participant the values within the significant region (t(13)=2.571, p=0.023, Cohen's d=0.687). The topography of the effect spread mostly over parietal and occipital regions (Fig. 4C), mimicking the topography of the EPs analysis. As previously, these results confirm the prediction that the SD task may involve additional computational mechanisms beyond feedforward computations, possibly indexed by the beta-band oscillatory processes identified here. As previously, these results confirmed those from the pilot experiment (figure 5D,E), confirming the robustness of the effect also in the oscillatory domain. Below, we contextualize and substantiate our results in light of the relevant literature.

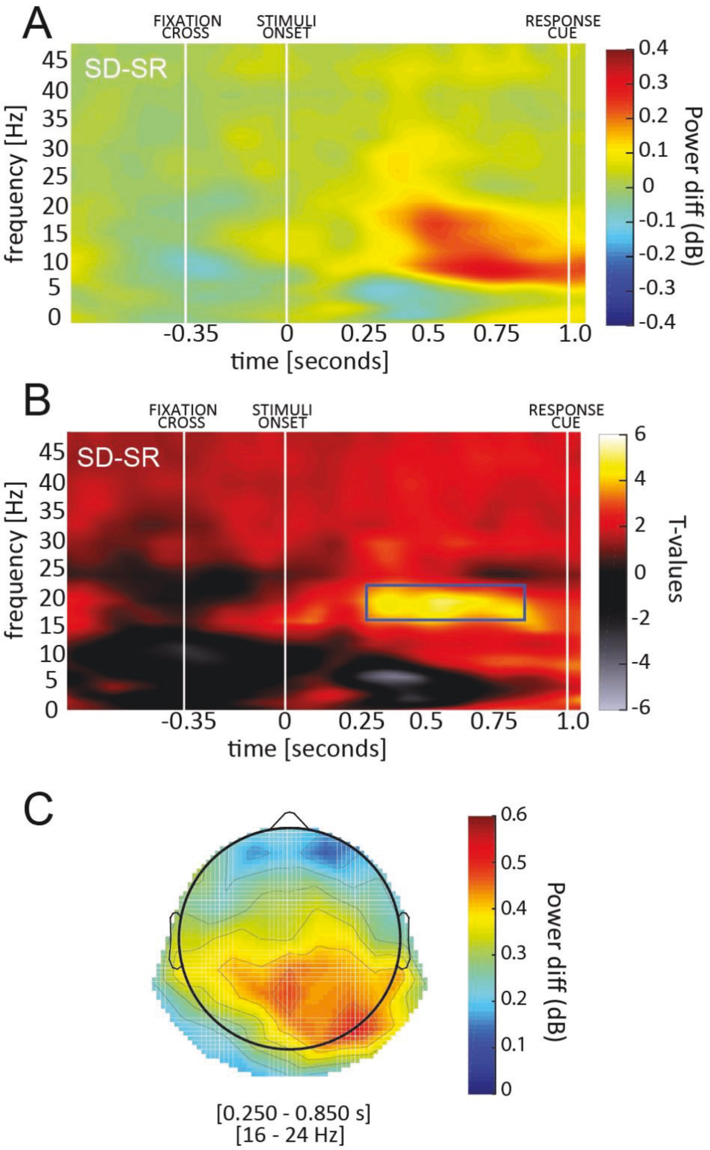

Fig.4: **Time-frequency results**. A) The difference between SD and SR power spectra is shown in the first panel. White lines indicate the onset of the fixation cross, the stimuli and the response cue. B) The second panel shows the corresponding t values (when testing the difference against zero). We observed a significant region in the low beta band (16-24Hz), between 250ms and 950ms after stimulus onset. C) The topography of the significant time-frequency window reveals the involvement of occipital-parietal regions.

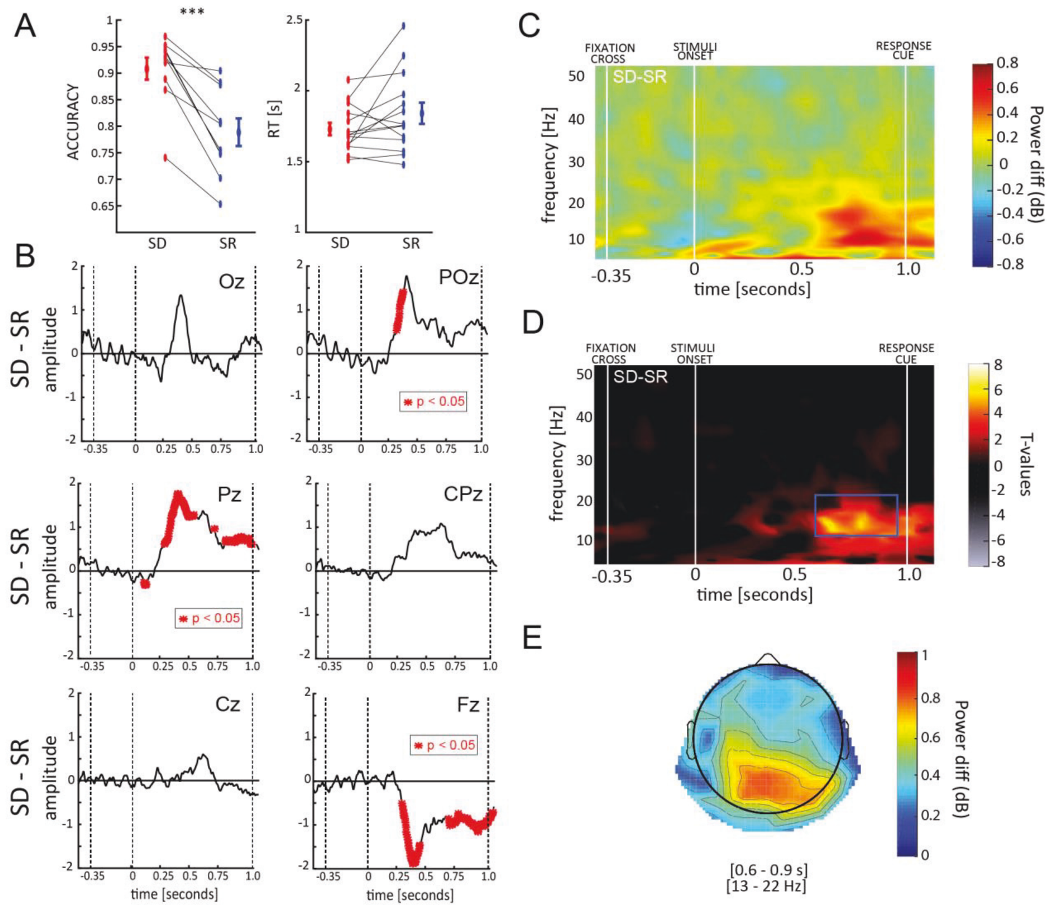

Fig.5: **Pilot experiment results**. A) Behavioral results of the pilot experiment: left and right panel shows accuracy and reaction times for SD (red) and SR (blue) tasks. Differently than in the main task, in the pilot experiment participants performed significantly better in the SD than in the SR task (compare the accuracy between figure 5A and 2B). B) Difference between SD and SR evoked potentials. Red asterisk indicate time window significantly different than zero C) Difference between SD and SR power spectra: white lines indicate the stimulus onset and the response cue. D) Testing the SD-SR difference against zero reveals a significant region in the low beta band (13-21Hz), before the response cue, in agreement with the results of the main experiment –figure 4. We reported a large effect size for this effect (one sample t-test against zero averaging

per each participant the values within the significant region, t(13)=7.049, p<0.001, Cohen’s d=1.820 E) As in the main experiment, the SD-SR difference mostly involves occipital-parietal regions.

4. Discussion

In this study, we confirmed in a series of two experiments a prediction from the computational study by Kim et al. (2018) that there exists an important dichotomy between visual reasoning tasks: While spatial relation (SR) tasks can be solved by modern deep convolutional neural networks (CNNs), same-different (SD) tasks pose a significant challenge for these architectures, suggesting the need for additional computations beyond a feedforward cascade of filtering, rectification and normalization operations. Importantly, the result of these simulations does not allow us to formulate any prediction about the specific cortical processes involved in the two tasks. Nonetheless, it demonstrates a fundamental computational difference, which can be tracked in terms of its human brain neural correlates while subjects solve SD vs. SR tasks (with difficulty objectively matched by an adaptive psychometric procedure). Remarkably, in both the pilot and the main experiment we found higher activity in the former task, in both evoked potentials and oscillatory components. We interpret these differences as reflecting additional computations required by the SD task. We can speculate that these additional computations involve working memory and attention processes, which are lacking in feedforward architectures such as CNNs.

Additionally, it is possible to interpret our results in a broader context, by considering other tasks supposed to involve spatial attention, such as visual search. Previous experimental work suggested the need for re-entrant processes (Treisman & Gelade, 1980; Wolfe, Cave, & Franzel, 1989), and how increased activity in specific oscillatory components (i.e. low [22-34Hz] and high [36-56Hz] gamma bands) are characteristic of these processes (Buschman & Miller, 2007; S. Phillips & Takeda, 2009). Accordingly, state-of-the-art computational models performing visual search and related tasks (e.g. instance segmentation) also employ attentional or recurrent mechanisms (Linsley, Ashok, Govindarajan, Liu, & Serre, 2020), supporting the hypothesis that convolutional feedforward networks can benefit from recurrent mechanisms in solving visual reasoning tasks (Kreiman & Serre, 2020).

Computational evidence for the hypothesis that the SD task requires additional computational mechanisms beyond those needed to solve the SR task is provided by the results of the Siamese network simulations (Bromley et al., 1994). This feedforward network processes each stimulus item in a separate (CNN) channel and then passes the processed items to a single classifier network. Since each item is processed separately (the network is fed two images with only one item represented in each), this ‘oracle’ architecture performs the task with item-segmentation processes automatically provided. Our results (as previously shown on another dataset by Kim and colleagues (Kim et al., 2018) demonstrate that such a feedforward network, once endowed with object individuation using the Siamese architecture, can easily learn to solve the SD task. In other words, this model simulates the beneficial effects of attentional selection, individuation and working memory by segregating the representations of each item. Our EEG results are compatible with this interpretation, with higher activity in the SD compared to the SR task, visible in both evoked potentials and oscillatory frequency bands that have been previously related to attention and working memory (Benchenane et al., 2011; Lundqvist et al., 2018; Nash & Fernandez, 1996).

Previous work has shown that modulations of beta-band oscillations can be related to selective attention mechanisms (Benchenane et al., 2011; Buschman & Miller, 2007; Lee, Whittington, & Kopell, 2013; Richter, Coppola, & Bressler, 2018). Different attentional mechanisms may indeed be involved in the two tasks: the SR task could be solved by first grouping items and then determining the orientation of the group (Franconeri, Scimeca, Roth, Helseth, & Kahn, 2012), whereas the SD task requires the individuation of the two items before comparison. In addition, our results are also consistent with differences in memory processes between the two tasks (de Fockert, G., Frith, & Lavie, 2001). One common assumption is that items that are grouped together (as in the SR task) occupy only one working memory slot (Clevenger & Hummel, 2014; Franconeri, Alvarez, & Cavanagh, 2013), whereas non-grouped items would each hold one slot, resulting in a larger working memory load. Previous literature showed that working memory can also be characterized by neuronal oscillatory signatures. Recent studies, for example, have demonstrated an interplay between beta and gamma band frequencies during working memory tasks (Lundqvist et al., 2016, 2018). Similarly, alpha and low beta bands, not only increase with working memory load (Babiloni et al., 2004; Pesonen, Hämäläinen, & Krause, 2007), but also in conjunction with the inhibition of competing visual memories in selective memory retrieval (Park, Min, & Lee, 2010; Waldhauser, Johansson, & Hanslmayr, 2012). Besides, previous studies have reported that increased oscillatory activity in the alpha band is a signature of attentional processes, and it can predict the likelihood of successful trials in many tasks (Händel, Haarmeier, & Jensen, 2011; Klimesch, 2012; Nelli, Itthipuripat, Srinivasan, & Serences, 2017); however, in our current study we did not investigate differences between correct and incorrect trials, but between different types of tasks (involving spatial relationship or sameness judgment), after controlling for task difficulty . This could explain why alpha-band amplitude differences were less prominent in our study. All considered, several lines of evidence point towards beta oscillations as crucially involved in both attention and working memory related processes. These processes, therefore, might be part of the additional computational mechanisms required for SD tasks compared to SR tasks. Future work could more directly compare the attention and memory dependence of each task in human subjects.

That feedforward neural networks are limited in their ability to solve simple visual reasoning tasks is already being recognized by the computer vision and neuroscience communities (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Rajalingham, Issa, Schmidt, Kar, & DiCarlo, 2017; Yamins, Hong, Cadieu, & Dicarlo, 2013). Current CNN extensions include modules for integrating local and global features (Chen et al., 2018) as well as recurrent neural architectures (Yang et al., 2018). Our results suggest that the human visual system also deploys additional computations beyond feedforward processes to successfully solve visual reasoning tasks. Rhythmic cortical oscillations in the beta-band represent the signatures of these additional computations, which may involve selective attention and working memory.

Acknowledgments

This work was funded by an ERC Consolidator Grant P-CYCLES number 614244 to RV, a joint CRCNS ANR-NSF Grant “OsciDeep” to RV (ANR-19-NEUC-0004) and TS (IIS-1912280), and two ANITI (Artificial and Natural Intelligence Toulouse Institute, ANR-19-PI3A-0004) Research Chairs to RV and TS. Additional support to TS was provided by ONR grant (N00014-19-1-2029).

References

Babiloni, C., Babiloni, F., Carducci, F., Cappa, S. F., Cincotti, F., Del Percio, C., … Rossini, P. M. (2004). Human cortical rhythms during visual delayed choice reaction time tasks: A high-resolution EEG study on normal aging. *Behavioural Brain Research*, *153*(1), 261–271. http://doi.org/10.1016/j.bbr.2003.12.012

Benchenane, K., Tiesinga, P. H., & Battaglia, F. P. (2011). Oscillations in the prefrontal cortex: A gateway to memory and attention. *Current Opinion in Neurobiology*. http://doi.org/10.1016/j.conb.2011.01.004

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436. http://doi.org/10.1163/156856897X00357

Bromley, J., Guyon, I., Lecun, Y., Sickinger, E., Shah, R., Bell, A., & Holmdel, L. (1994). Signature Verification using a "Siamese" Time Delay Neural Network. In *Advances in neural information processing systems* (pp. 737--744). Retrieved from http://papers.nips.cc/paper/769-signature-verification-using-a-siamese-time-delay-neural-network.pdf

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*(5820), 1860–1864. http://doi.org/10.1126/science.1138071

Chen, X., Li, L.-J., Fei-Fei, L., & Gupta, A. (2018). Iterative Visual Reasoning Beyond Convolutions. http://doi.org/10.1109/CVPR.2018.00756

Clevenger, P. E., & Hummel, J. E. (2014). Working memory for relations among objects. *Attention, Perception, and Psychophysics*, *76*(7), 1933–1953. http://doi.org/10.3758/s13414-013-0601-3

Daniel, T. A., Wright, A. A., & Katz, J. S. (2015). Abstract-concept learning of difference in pigeons. *Animal Cognition*, *18*(4), 831–837. http://doi.org/10.1007/s10071-015-0849-1

de Fockert, J. W., G., R., Frith, C. D., & Lavie, N. (2001). The role of working memory in visual selective attention . *S*, *291*, 1803–1806.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. http://doi.org/10.1016/j.jneumeth.2003.10.009

Fabiani, M., Karis, D., & Donchin, E. (1986). P300 and Recall in an Incidental Memory Paradigm. *Psychophysiology*, *23*(3), 298–308. http://doi.org/10.1111/j.1469-8986.1986.tb00636.x

Franconeri, S. L., Alvarez, G. A., & Cavanagh, P. (2013). Flexible cognitive resources: Competitive content maps for attention and memory. *Trends in Cognitive Sciences*. http://doi.org/10.1016/j.tics.2013.01.010

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227. http://doi.org/10.1016/j.cognition.2011.11.002

GoogleResearch. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. *Google Research*. http://doi.org/10.1207/s15326985ep4001

Händel, B. F., Haarmeier, T., & Jensen, O. (2011). Alpha oscillations correlate with the successful inhibition of unattended stimuli. *Journal of Cognitive Neuroscience*, *23*(9), 2494–2502. http://doi.org/10.1162/jocn.2010.21557

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2016–Decem, pp. 770–778). http://doi.org/10.1109/CVPR.2016.90

Hochberg, B. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. http://doi.org/10.2307/2346101

Hollard, V. D., & Delius, J. D. (1982). Rotational invariance in visual pattern recognition by pigeons and humans. *Science*, *218*(4574), 804–806. http://doi.org/10.1126/science.7134976

Itthipuripat, S., Cha, K., Byers, A., & Serences, J. T. (2017). Two different mechanisms support selective attention at different phases of training. *PLoS Biology*, *15*(6). http://doi.org/10.1371/journal.pbio.2001724

Itthipuripat, S., Cha, K., Deering, S., Salazar, A. M., & Serences, J. T. (2018). Having more choices changes how human observers weight stable sensory evidence. *Journal of Neuroscience*, *38*(40), 8635–8649. http://doi.org/10.1523/JNEUROSCI.0440-18.2018

JASP Team. (2018). JASP (Version 0.8.6.0). *[Computer Software]*. Retrieved from http://jasp-stats.org

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983. http://doi.org/10.1038/s41593-019-0392-5

Kim, J., Ricci, M., & Serre, T. (2018). Not-So-CLEVR: Learning same-different relations strains feedforward neural networks. *Interface Focus*, *8*(4). http://doi.org/10.1098/rsfs.2018.0011

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic gradient descent. *ICLR: International Conference on Learning Representations*.

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*. http://doi.org/10.1016/j.tics.2012.10.007

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*(3), 557–577. http://doi.org/10.1017/S0048577201990559

Kreiman, G., & Serre, T. (2020). Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*. http://doi.org/10.1111/nyas.14320

Krusemark, E. A., Kiehl, K. A., & Newman, J. P. (2016). Endogenous attention modulates early selective attention in psychopathy: An ERP investigation. *Cognitive, Affective and Behavioral Neuroscience*, *16*(5), 779–788. http://doi.org/10.3758/s13415-016-0430-7

Lee, J. H., Whittington, M. A., & Kopell, N. J. (2013). Top-Down Beta Rhythms Support Selective Attention via Interlaminar Interaction: A Model. *PLoS Computational Biology*, *9*(8). http://doi.org/10.1371/journal.pcbi.1003164

Linsley, D., Ashok, A. K., Govindarajan, L. N., Liu, R., & Serre, T. (2020). Stable and expressive recurrent vision models. Retrieved from http://arxiv.org/abs/2005.11362

Love, J., Selker, R., Verhagen, J., Marsman, M., Gronau, Q. F., Jamil, T., … Rouder, J. N. (2015). Software to sharpen your stats. *APS Observer*, *28*(3), 27–29.

Lundqvist, M., Herman, P., Warden, M. R., Brincat, S. L., & Miller, E. K. (2018). Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature Communications*, *9*(1). http://doi.org/10.1038/s41467-017-02791-8

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L. L., Buschman, T. J. J., & Miller, E. K. K. (2016). Gamma and Beta Bursts Underlie Working Memory. *Neuron*, *90*(1), 152–164. http://doi.org/10.1016/j.neuron.2016.02.028

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. http://doi.org/10.1016/j.jneumeth.2007.03.024

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing, 679–690. http://doi.org/10.3758/s13428-010-0049-5

McEvoy, L. (1998). Dynamic cortical networks of verbal and spatial working memory: effects of memory load and task practice. *Cerebral Cortex*, *8*(7), 563–574. http://doi.org/10.1093/cercor/8.7.563

Nash, A. J., & Fernandez, M. (1996). P300 and allocation of attention in dual-tasks. *International Journal of Psychophysiology*, *23*(3), 171–180. http://doi.org/10.1016/S0167-8760(96)00049-9

Nelli, S., Itthipuripat, S., Srinivasan, R., & Serences, J. T. (2017). Fluctuations in instantaneous frequency predict alpha amplitude during visual perception. *Nature Communications*, *8*(1). http://doi.org/10.1038/s41467-017-02176-x

Park, H. D., Min, B. K., & Lee, K. M. (2010). EEG oscillations reflect visual short-term memory processes for the change detection in human faces. *NeuroImage*, *53*(2), 629–637. http://doi.org/10.1016/j.neuroimage.2010.06.057

Pesonen, M., Hämäläinen, H., & Krause, C. M. (2007). Brain oscillatory 4-30 Hz responses during a visual n-back memory task with varying memory load. *Brain Research*, *1138*(1), 171–177. http://doi.org/10.1016/j.brainres.2006.12.076

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., … O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6171–6176. http://doi.org/10.1073/pnas.1721355115

Phillips, S., & Takeda, Y. (2009). Greater frontal-parietal synchrony at low gamma-band frequencies for inefficient than efficient visual search in human EEG. *International Journal of Psychophysiology*, *73*(3), 350–354. http://doi.org/10.1016/j.ijpsycho.2009.05.011

Rajalingham, R., Issa, E. B., Schmidt, K., Kar, K., & DiCarlo, J. J. (2017). Feedforward deep neural networks diverge from humans and monkeys on core visual object recognition behavior. In *Annual Conference on Cognitive Computational Neuroscience* (pp. 1–2). Retrieved from https://www2.securecms.com/CCNeuro/docs-0/59288a3f68ed3f3c458a257f.pdf

Richter, C. G., Coppola, R., & Bressler, S. L. (2018). Top-down beta oscillatory signaling conveys behavioral context in early visual cortex. *Scientific Reports*, *8*(1). http://doi.org/10.1038/s41598-018-25267-1

Serre, T. (2019). Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, *5*(1), 399–426. http://doi.org/10.1146/annurev-vision-091718-014951

Smith, J. M. B. and A. F. M. (2001). Bayesian Theory. *Measurement Science and Technology*. http://doi.org/10.1088/0957-0233/12/2/702

Stabinger, S., Rodríguez-Sánchez, A., & Piater, J. (2016). 25 years of CNNS: Can we compare to human abstraction capabilities? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9887 LNCS, pp. 380–387). http://doi.org/10.1007/978-3-319-44781-0_45

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. Retrieved from http://www.erc.caltech.edu/Industry/Conferences/2004/AIC/pdf/Laurent-Itti.pdf

Van Voorhis, S., & Hillyard, S. A. (1977). Visual evoked potentials and selective attention to points in space. *Perception & Psychophysics*, *22*(1), 54–62. http://doi.org/10.3758/BF03206080

VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, *30*(6), 655–668. http://doi.org/10.1068/p3029

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 1: behavioural study. *European Journal of Neuroscience*, *11*(4), 1223–1238. Retrieved from http://onlinelibrary.wiley.com/doi/10.1046/j.1460-9568.1999.00530.x/full%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/10103118

Waldhauser, G. T., Johansson, M., & Hanslmayr, S. (2012). Alpha/Beta Oscillations Indicate Inhibition of Interfering Visual Memories. *Journal of Neuroscience*, *32*(6), 1953–1961. http://doi.org/10.1523/jneurosci.4201-11.2012

Wasserman, E. A., Castro, L., & Freeman, J. H. (2012). Same-different categorization in rats. *Learning and Memory*, *19*(4), 142–145. http://doi.org/10.1101/lm.025437.111

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120. http://doi.org/10.3758/BF03202828

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An Alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419–433. http://doi.org/10.1037/0096-1523.15.3.419

Yamins, D., Hong, H., Cadieu, C., & Dicarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to Macaque IT and human ventral stream. In *Advances in Neural Information Processing Systems*.

Yang, G. R., Ganichev, I., Wang, X. J., Shlens, J., & Sussillo, D. (2018). A dataset and architecture for visual reasoning with a working memory. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 11214 LNCS, pp. 729–745). http://doi.org/10.1007/978-3-030-01249-6_44