# Coherence of Working Memory Study Between Deep Neural Network and Neurophysiology

Yurui Ming, Chin-Teng Lin

***Abstract*—The automatic feature extraction capability of deep neural networks (DNNs) endows them with the potentiality for analyzing complicated electroencephalogram (EEG) data captured from brain functionality research. This work investigates the potential coherent correspondence between the region-of-interest (ROI) for DNN to explore, and ROI for conventional neurophysiological-oriented methods to work with, as exemplified in the case of working memory study. The attention mechanism induced by the global average pooling (GAP) is applied to a public EEG dataset of working memory test, to unveil these coherent ROIs via a classification problem. The results show the potential alignment of the ROIs from different discipline methods, and consequently asserts the confidence and promise of utilizing DNN for EEG data analysis.**

***Index Terms*—Attention Mechanism, Deep Neural Network (DNN), Electroencephalogram (EEG), Working Memory (WM).**

## I. Introduction

The success of deep neural networks (DNNs) in various fields has drawn the attention of brain researchers to apply these models for electroencephalogram (EEG) data analysis, either to promote deeper neuroscientific understandings or to facilitate wider brain-computer interface (BCI) applications [1-4]. Although it is not as strict as clinical requirement, the black-box operations of DNNs still arouse lots of concerns. For example, it is difficult to reach intuitive interpretations to the model behavior without knowing the underlying mechanism. Despite ways to interpret the neural network dynamics to foster the intuition [5-7], approaches to combine techniques in other disciplines to improve the performance [8-11], and methods to quantify the uncertainty of DNNs to increase the trustworthiness [12], it is still necessary to study the characteristics and to assert the feasibility of network models by linking and comparing the achievements from other methodologies.

The properties such as high temporal resolution, mobility, economy, etc. [13], confer EEG's indispensable role in brain research. To blindly employ DNNs to analyze EEG data might provide satisfying results catered to the application itself; however, loss of intuition can hinder the theoretical depth of the achievements. Compared with other DNN-affiliated applications, certain brain research experiments are only conducted in restricted environments or idealized conditions due to practical constraints, and the results are then demonstrated to the community without actual deployments of the models. This can lead to the potential unintended consequences being buried in the learned models when using neural networks, thus turning the overall research into hallucinations and can also produce confusing results at a later stage [14-16]. Hence, to cross validate the results of DNN models in EEG data analysis by referencing knowledge from other disciplines is more critical in this field than other utilizations of DNNs.

This work aims to address part of these concerns by considering the implicit attention mechanism induced by class activation mapping via global average pooling (GAP) [17]. Actually, for certain fundamental brain research topics, such as working memory [18], there are already common recognitions of the neuronal basis underpinning this mechanism. It is regarded that the prefrontal cortex (PFC) and hippocampus are actively involved in the functioning of working memory [19-21]. Therefore, it is expected that for the working memory load test, when a DNN is applied for harvesting EEG features automatically [2], the model should not switch absurdly among different areas over the scalp. Although there is still some dispute over the characteristics of neuronal activities, such as in essence these are discrete dynamics or sustained activities [22], the distinction among workload extents should only incur the network to focus on approximately common areas.

Work in [17] demonstrates that by adoption of GAP, even if trained via class level labels, the network can still exhibits some localization ability, which is able to identify the discriminative image regions in a picture. This apparent simplicity provides the means to verify some conjectures, for example, in the case of working memory, whether the network tends to look at similar areas with different activation strengths or not. It is known that due to the still limited understanding of working memory, enforcing network to explicitly explore specific regions might not be a good idea. Instead, the network's intrinsic dynamics and behaviors, if they are in accordance with certain assumptions, are strong evidence to the feasibility of utilizing DNNs to analyze EEG data. And this is the purpose and potential contribution of our work in this paper.

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

Yurui Ming and Chin-Teng Lin are with the School of Computer Science, Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW 2007 Australia (e-mail: yrming@gmail.com, Chin-Teng.Lin@uts.edu.au).

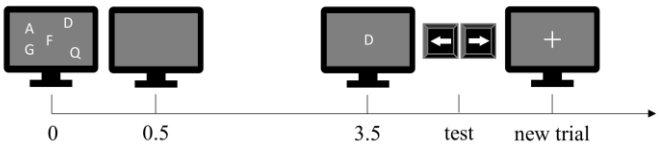


Fig. 1 Experiment paradigm for conducting the research.

## II. METHODOLOGY

As mentioned above, working memory is an indispensable component in many important cognitive functions and processes. Irrespective of whether the elicited spiking pattern of neurons is continuously sustained or discretely dynamic, it is believed that high-order cortical areas, such as the PFC, are highly involved [23, 24]. Therefore, it is interesting to check where the network focus (the region-of-interest or ROI) is capable of distinguishing among the workload extent or not during different workloads exerted on the working memory.

### A. Experiment and Data

Fig. 1 shows the overall setup of the experiment, which is used to investigate the behavior of a given network employed to analyze the EEG data captured during a working memory capacity test [25]. The paradigm of the working memory task for EEG dataset acquisition is as follows. A randomized letter set is displayed for the participant at 0 second (s) relative to the beginning of current trial. It lasts for 0.5 s then fades away. At 3.5 s, a letter appears at the screen and subject needs to decide whether it belongs to the previous shown letter set or not on perceiving the letter. After the subject's action, a fixation icon launches a new trial.

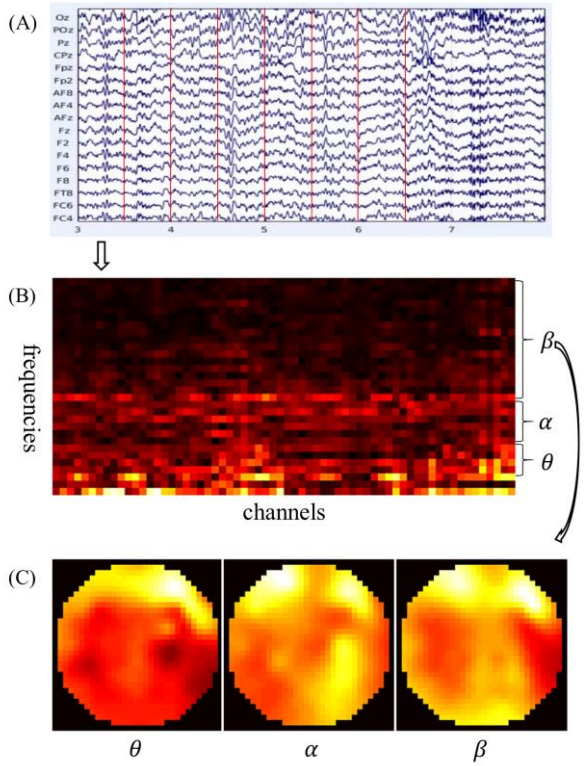


Fig. 2 Preparations of topographical EEG data in the spatial domain from waveform EEG data in the time domain.

The EEG signals are consecutively captured during the entire session which comprises various trials, and the data preparation is the same as in [2]. In detail, the EEG time series data lasting for 3.5 s are sliced into 7 non-overlapping segments. Then Fourier transform is applied to each segment to obtain the power spectrum up to 30Hz for each channel. According to the literature which reveals the effectiveness of using respective EEG sub-bands [25, 26], three bands, i.e., theta (4 – 7 Hz), alpha (8 – 13 Hz), and beta (14 – 30 Hz) are considered in this research. For each band of the individual channel, the squared absolute values within the frequency band are added up to measure the contribution of the electrode source. Finally, together with the corresponding EEG montage used in the experiment, the topographical representations are generated. Fig. 2 shows the overall procedures in an illustrative manner. Fig. 2(A) displays the EEG waveform data in the time domain; Fig. 2(B) shows each segment converted into the frequency domain by Fast Fourier Transform (FFT). The $x$-axis indicates the channels and the $y$-axis represents the absolute values of the frequency components; Fig. (C) presents the EEG topography generated from the frequency data by interpolation and extrapolation according to the coordinates of the electrode placement.

### B. Network Architecture

The corresponding constructed DNN for analyzing EEG data to unveil the network dynamics is in Fig. 3. First, a convolutional network (convnet) is applied to the non-overlapping segmented topographical data at each time step. Note the weights of the convnet are shared when processing each segment. Then the processed data are fed into a recurrent network for further computation. To better explore the spatial information of the topographies, convolution is used inside the recurrent cell instead of conventional linear transformation plus non-linear activation. Hence, the subnetwork is abbreviated as recur-convnet. In detail, assuming at time step $t$, the input to the recur-convnet is $x_t$, the cell state is $c_t$, the hidden state is $h_t$, and the values of the input gate, forget gate and output gate are $i_t$, $f_t$ and $o_t$, respectively. The computation is governed by the following formulas:

$$\chi_t = \mathrm{concat}(i_t, h_{t-1}) \tag{1}$$

$$[i_t, f_t, o_t] = \mathrm{conv}(\chi_t, w_g) \tag{2}$$

$$c_t = \sigma(f_t) * c_{t-1} + \sigma(i_t) * \tan(\mathrm{conv}(\chi_t, w_i)) \tag{3}$$

$$h_t = \sigma(o_t) * \tan(c_t) \tag{4}$$

where $w$ represents the weights of different network circuits in above formulas.

After the recurrence, only feature maps of the last step are considered. GAP is then applied to these feature maps to obtain resulting weights before the final classification. These feature maps and weights are used to construct the heatmaps as in [17].

The detailed configuration of the network structure is provided in TABLE I.

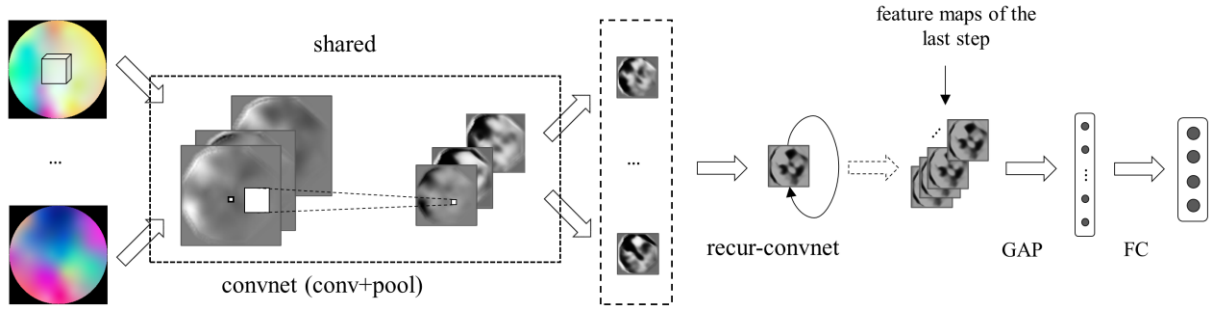


Fig. 3 Architecture of the neural network for analyzing EEG data.

TABLE I
NETWORK ARCHITECTURE

| Block | Layer | Filters | Size | Activation | Padding |
|---|---|---|---|---|---|
| ConvNet | Conv2D | 8 | (3, 3) | ReLU | SAME |
| | Conv2D | 8 | (3, 3) | ReLU | SAME |
| | Conv2D | 8 | (3, 3) | ReLU | SAME |
| | Conv2D | 8 | (3, 3) | ReLU | SAME |
| | AvgPool | | (2, 2) | | |
| | Conv2D | 16 | (3, 3) | ReLU | SAME |
| | Conv2D | 16 | (3, 3) | ReLU | SAME |
| | AvgPool | | (2, 2) | | |
| | Conv2D | 32 | (3, 3) | ReLU | SAME |
| | AvgPool | | (2, 2) | | |
| Recur-ConvNet* | Conv2D | 128 | (3, 3) | | SAME |
| GAP | AvgPool | | (8, 8) | | VALID |
| FC | Linear | | #class | Softmax | |

*Number of recurrences is 7

### C. Heatmap Generation and Investigation

The complexity of the cognition and data acquisition process brings in the non-intuition to the understanding of the EEG data, which consequently hinders the interpretation of the network outcome. To explicitly demonstrate the implicit attention capability of the network, a natural dog-vs-cat image set [27], which is compiled for categorizing dogs and cats, is used to verify the ROIs of the network when performing classification in the first place. Because the designed network has a recurrent part, for each dog or cat image, it repeats for 7 times to form a sequential sample to resemble the EEG topographical data. This means the layouts of the respective input data are identical between two datasets, just the images from the dog-vs-cat set are natural and straightforward for observation and evaluation.

Therefore, the designed network is first trained with the dog-vs-cat images to decide the hyperparameters. After training, the test images are fed into the network for prediction and simultaneously to obtain the heatmaps. For heatmap generation, the weighted summation of all the feature-maps from the last step of the recur-convnet is to form the index matrix, which is used to index into the values of a chosen colormap. The weights involved in the summation are the ones accompanying the post-GAP layer. Representatives of these achieved intuitive heatmaps which prove the effectiveness of the designed network are illustrated in Fig. 4.

As the next step, the prepared EEG topographical data are processed by the network. A leave-one-subject-out test paradigm is considered here, which means that each time the data of one subject is fixed and the data of all the remaining subjects are used to train the network. The hyperparameters are directly migrated from the case of the training on the dog-vs-cat dataset. Since based on these hyperparameters, the implicit attention induced by the network can be intuitively observed, it is conjectured that the consequent training with the same network architecture on EEG topographical dataset could unveil the implicit attention as well.

We compare the test accuracies achieved via the network in this paper with the state-of-the-art (SOTA) results obtained by a recursive convolutional neural network (RCNN) in TABLE II. Because the network here might not be an optimized one to target the mind load classification, and GAP which collapses the whole feature-map into one point is still too coarse, the obtained test accuracy in this paper is not as good as the SOTA results. However, the purpose in this paper is to demonstrate the attention mechanism is an intrinsic property of a rather general neural network, in this regard, a fairly comparable result is also

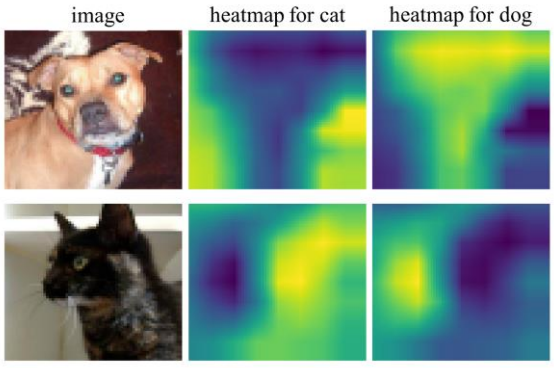


Fig. 4 Implicit attention capability of the designed network verified by the dog-vs-cat image set.

TABLE II
Accuracies of Memory Workload Classification for Each Subject

| Test Subject | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RCNN* | 88.9 | 76.5 | 93.3 | 99 | 100 | 98 | 100 | 98.5 | 99 | 96.8 | 96.5 | 91 | 46.8 | 91.1 |
| GAP | 57.3 | 64.6 | 84.9 | 87.6 | 96.4 | 94 | 96.4 | 97.5 | 91.9 | 88.4 | 82.9 | 67 | 47.3 | 81.3 |

*Statistics is directly from [2]

acceptable.

For heatmap generation, to minimize the wrong predictions which can interfere the following investigation, only samples with the right predictions are considered. The above procedure is repeated for all subjects with correctly predicted samples and the corresponding heatmaps counting 2174 for analysis.

To investigate the properties of these heatmaps, for example, whether heatmaps corresponding to samples under different mind workload are separated or not, t-SNE is utilized to observe the distribution of these heatmaps in lower dimensional (D) spaces, such as in two-dimensional space [28]. Because the heatmaps are of high-dimensional data potentially spreading along certain manifold, instead of using Euclidean distance, the structural similarity index (SSIM) is considered here [29]. As inferred from [29], the SSIM measures the perceptual difference between two similar images and cannot judge which of the two is better. This is permittable with this research, since whether a specific heatmap is good or not is not confirmative, and only the collective heatmaps can propose certain conclusions. Therefore, as in [28], for heatmaps $h_i$ and $h_j$, the conditional probability is calculated as in (5):

$$p_{j|i} = \frac{\exp\left(\beta\left(1-\mathrm{ssim}\left(h_i,h_j\right)\right)^2\right)}{\sum_{j\neq k}\exp\left(\beta\left(1-\mathrm{ssim}\left(h_i,h_k\right)\right)^2\right)} \tag{5}$$

The perplexity of t-SNE is set at 20 to map the heatmaps from a high dimensional space of 4096 dimensions into a 2-D plane. The algorithm is iterated for 1000 times to obtain the distribution of heatmaps in low dimension for examination.

## III. Results

Fig. 5 illustrates the t-SNE result, which can help to inspect the clustering attribute of all the heatmaps. It is manifest that each class displays certain region preferences, while meantime being spread over the overall plane. However, no class obviously dominates any specific region that clearly separates itself from other classes. This to some extent indicates that the network might work on some common parts of EEG topographies to distinguish the workload. These coherent parts of EEG topographies in turn indicate the activation of the corresponding brain regions. Fig. 5 indicates that for each class, the network will not focus differently, at least apparently, which is optimistically supporting our conjecture. In addition, in Fig. 5, the clustering of class 2 is a little different from the other classes. An explanation could be articulated as follows. For class 1, the cognition required to process the workload might be too simple to be distinguishable from the instinct; for class 4, the workload could be too high and complicated that requires sophisticated cognitive functions to be involved. Class 2 which represents an appropriate workload is probably suitable for arousing the cognition of working memory. This might explain its distinctiveness.

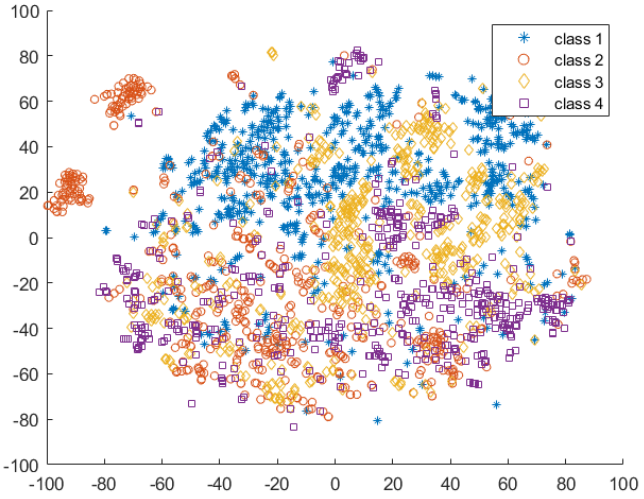


Fig. 5 Heatmap distribution under different workloads via t-SNE.

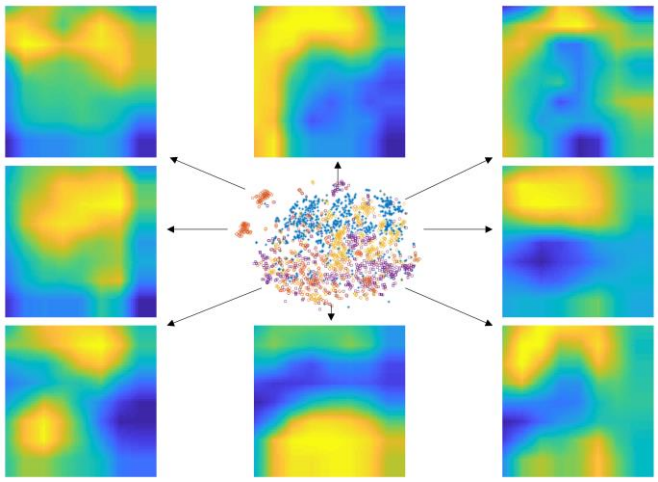

Fig. 6 Heatmaps from different locations of the t-SNE distribution.

To further investigate the involved brain region, several heatmaps from different locations in Fig. 5 are displayed in Fig. 6 to highlight this. It is interesting to notice that from these heatmaps, the PFC gets highly focused in almost all cases, although the activations might be to varying degrees. However, the result is sufficient to demonstrate the accordance with the conclusion of the neurophysiological study mentioned in the introduction part.

## IV. Conclusion

This paper investigated the potential conclusion coherence of working memory study between the approach from the DNN-based perspective and the methods from the neurophysiological perspective. The results revealed the fact that the network overall looks into certain common areas to distinguish the topographical EEG data under different workloads. Also, the brain areas tend to get focused are in the PFC area, which merits the conclusions made from the conventional neural physiological studies.

## References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015/05/01 2015, doi: 10.1038/nature14539.

[2] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks," arXiv.org, 2016.

[3] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces," (in eng), J Neural Eng, vol. 15, no. 5, p. 056013, Oct 2018, doi: 10.1088/1741-2552/aace8c.

[4] A. Vahid, M. Mückschel, S. Stober, A.-K. Stock, and C. Beste, "Applying deep learning to single-trial EEG data provides evidence for complementary theories on action control," Communications Biology, vol. 3, no. 1, p. 112, 2020/03/09 2020, doi: 10.1038/s42003-020-0846-z.

[5] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, 2014: Springer, pp. 818-833.

[6] J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability," arXiv.org, 2017.

[7] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," Digital signal processing, vol. 73, no. C, pp. 1-15, 2018, doi: 10.1016/j.dsp.2017.10.011.

[8] K. Vinken, X. Boix, and G. Kreiman, "Incorporating intrinsic suppression in deep neural networks captures dynamics of adaptation in neurophysiology and perception," Science advances, vol. 6, no. 42, p. eabd4205, 2020.

[9] W. Fruehwirt et al., "Bayesian deep neural networks for low-cost neurophysiological markers of Alzheimer's disease severity," arXiv preprint arXiv:1812.04994, 2018.

[10] S. M. J. Jalali, S. Ahmadian, A. Khosravi, M. Shafie-khah, S. Nahavandi, and J. P. Catalao, "A Novel Evolutionary-based Deep Convolutional Neural Network Model for Intelligent Load Forecasting," IEEE Transactions on Industrial Informatics, 2021.

[11] S. M. J. Jalali et al., "Towards novel deep neuroevolution models: chaotic levy grasshopper optimization for short-term wind speed forecasting," Engineering with Computers, pp. 1-25, 2021.

[12] M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," arXiv preprint arXiv:2011.06225, 2020.

[13] Niedermeyer's electroencephalography : basic principles, clinical applications, and related fields, Sixth edition. ed. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins Health, 2011.

[14] J. Shane. "When algorithms surprise us." https://aiweirdness.com/post/172894792687/when-algorithms-surprise-us (accessed 01/10, 2020).

[15] M. Minsky. "Embarrassing mistakes in perceptron research." https://www.webofstories.com/play/marvin.minsky/122 (accessed 01/10, 2020).

[16] M. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," arXiv.org, 2016.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016 2016, pp. 2921-2929, doi: 10.1109/CVPR.2016.319.

[18] B. Alan, "Working memory: looking back and looking forward," Nature Reviews Neuroscience, vol. 4, no. 10, p. 829, 2003, doi: 10.1038/nrn1201.

[19] J. O. Pernille, W. Helena, and K. Torkel, "Increased prefrontal and parietal activity after training of working memory," Nature Neuroscience, vol. 7, no. 1, p. 75, 2003, doi: 10.1038/nn1165.

[20] J. D. Murray, A. Bernacchia, N. A. Roy, C. Constantinidis, R. Romo, and X.-J. Wang, "Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex.(Report)," Proceedings of the National Academy of Sciences of the United States, vol. 114, no. 2, p. 394, 2017, doi: 10.1073/pnas.1619449114.

[21] E. Boran et al., "Persistent hippocampal neural firing and hippocampal-cortical coupling predict verbal working memory load," Science Advances, vol. 5, no. 3, p. eaav3687, 2019, doi: 10.1126/sciadv.aav3687.

[22] M. Lundqvist, J. Rose, P. Herman, Scott l. Brincat, Timothy j. Buschman, and Earl k. Miller, "Gamma and Beta Bursts Underlie Working Memory," Neuron (Cambridge, Mass.), vol. 90, no. 1, pp. 152-164, 2016, doi: 10.1016/j.neuron.2016.02.028.

[23] E. K. Miller, C. A. Erickson, and R. Desimone, "Neural mechanisms of visual working memory in prefrontal cortex of the macaque," The Journal of neuroscience : the official journal of the Society for Neuroscience, vol. 16, no. 16, pp. 5154-5167, 1996, doi: 10.1523/JNEUROSCI.16-16-05154.1996.

[24] M. G. Stokes, "'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework," Trends in Cognitive Sciences, vol. 19, no. 7, p. 394, 2015.

[25] P. Bashivan, G. M. Bidelman, and M. Yeasin, "Spectrotemporal dynamics of the EEG during working memory encoding and maintenance predicts individual behavioral capacity," European Journal of Neuroscience, vol. 40, no. 12, pp. 3774-3784, 2014, doi: 10.1111/ejn.12749.

[26] O. Jensen and C. D. Tesche, "Frontal theta activity in humans increases with memory load in a working memory task," European Journal of Neuroscience, vol. 15, no. 8, pp. 1395-1399, 2002, doi: 10.1046/j.1460-9568.2002.01975.x.

[27] M. Research. "Dogs vs. Cats Create an algorithm to distinguish dogs from cats." https://www.kaggle.com/c/dogs-vs-cats/data (accessed 01/12, 2020).

[28] L. J. P. van der Maaten and G. E. Hinton, "Visualizing High-Dimensional Data Using t-SNE," Journal of machine learning research, vol. 9, no. nov, pp. 2579-2605, 2008.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600-612, 2004.