VISIBLE COALITIONS OF NEURONAL ACTIVITIES IN BRAIN-TO-TEXT COMMUNICATION VIA HANDWRITING

John Bolognino[1]*, Sarel Cohen*[2]*, Eden Bar*[2]*, Rimon Shushan* [2]*, Patrisia Kaplun*[2]*,*
Dor Katirachi[2], Dvir Cohen[2], George Kour[3], Peter Chin[4], Eilon Vaadia[5]

[1]Boston University, Boston MA 02115, jcbolo@bu.edu
[2]The Academic College of Tel Aviv-Yaffo, Israel
sarelco@mta.ac.il, bared@mta.ac.il, rimonsh@mta.ac.il,
patrisiaka@mta.ac.il, dorki@mta.ac.il, dvirch@mta.ac.il
[3]IBM Research, Israel gkour@ibm.com
[4]Dartmouth College, Hanover, NH 03755, peter.chin@dartmouth.edu
[5]The Edmond and Lily Safra Center for Brain Sciences at
The Hebrew University of Jerusalem, Israel, eilon.vaadia@elsc.huji.ac.il

ABSTRACT

The brain's cortex features complex networks composed of many individual neurons [1, 2]. Various studies have revealed that the connectivity among neurons may vary in relation to behavioral events [3, 4, 5, 6, 7]. In a recent study, Willett *et al.* [8] demonstrated decoding of imagined handwriting movements from neural activity in the motor cortex of a paralyzed patient. We analyzed their data[1] by representing all neurons as raster displays and trained convolutional neural network (CNN) models to classify different brain states as the characters that the subjects imagined. Our binary classification models had an average accuracy of 96%, which we then fine-grained by training a multi-class CNN on all 31 different characters. This achieved a high success rate of 86% accuracy. Finally, we applied Grad-CAM [9] to explore the emergence of spatiotemporal patterns which are likely to be involved in determining which character the subject was imagining. Our results support the notion that dynamic neuronal correlations are involved in encoding the different characters.

Index Terms— Computational Neuroscience, Brain-computer-interface (BCI), Neuronal Dynamics, Brain-to-Text (BTT), Image Classification, Grad-CAM, Explainable AI

1. INTRODUCTION

Individual neurons in the cortex form intricate local networks [1, 2]. Numerous studies have revealed dynamic shifts in functional connectivity among neurons. [3, 4, 5, 6, 7]. The emergence of large-scale recording methods over the past two decades has reshaped our perspective of neuronal coding, from the single unit doctrine [10] to the notion that neural networks [11, 12, 13] with dynamic correlation structure, constitute the mechanisms of computation and representations in the brain [14, 15, 16, 17]. Numerous studies have turned to multidimensional analysis of population activity within the neuronal space. This conceptual space employs each dimension to represent the activity of an individual neuron, and at every time point, a set of neurons can be symbolized by a solitary point within a neuronal space. The positioning of this point is determined by the activity level of the corresponding neuron for each dimension [18, 19, 20, 21].

Here, we applied a CNN to data recorded by [8] while a human subject was engaged in imagining writing English characters (see details of the experiments in [8]). We aimed to achieve optimal accuracy in detecting a target character by the neuronal activity at a given time interval. Furthermore, we attempted to examine the notion that neural coding in the brain is based on the dynamic activity of groups of neurons, by using the explainable AI technique of Gradient-weighted Class Activation Mapping (Grad-CAM) [9] to unveil these visual patterns of neuronal activities.

While our deep learning techniques are well known, we innovate in our unique application of CNN and explainable AI techniques in such a way that the activations of our artificial neural network elucidate the activation patterns of neurons in the human brain, particularly for the task of handwriting and provide an additional visual manifestation computation by cortical dynamical networks.

2. METHODOLGY

The main analyses of this work used an artificial neural network (CNN) that learns features from input images, using layers of filters to identify patterns in spatially adjacent pixels.

[1]**Acknowledgment:** we thank Willett *at. el.* [8] for making their dataset publicly available.

While previous work has made use of Recurrent Neural Networks (RNNs), which are useful for encoding and performing classification on temporal data, CNNs can extract spatial data. For a dataset comprised of raster plots, which are two-dimensional matrices with electrodes (neuronal spikes) on the vertical axis and time-bins on the horizontal axis. A CNN can make predictions based on both temporal and spatial cues. The linear arrangement of the electrodes on the y-axis of the raster plots determines a spatial order to the data, within which the CNN can identify patterns of correlated activation. While the physical electrodes are arranged in a grid formation (Utah Arrays, 4X4mm grid), the raster plot shows their activity such that each row is placed end-to-end in a vertical line on the y-axis. This formation results in close spatial proximity of rows of electrodes in the recording which are physically distant from one another in the brain, providing our model with both nearby and distant electrode relationships to learn patterns from.

More importantly, CNNs have the benefit of being explainable. Since CNNs are intended to work with two-dimensional visual data, a variety of tools exist to demonstrate how and why trained CNN models make the predictions that they do. We utilized Gradient-weighted Class Activation Mapping (Grad-CAM), to identify groups of electrodes whose correlated activity informed our model's predictions.

We began by training a binary classifier using a ResNet50 CNN with binary cross-entropy loss to determine how well it could distinguish raster plots representing one character versus plots representing all other characters, (for example 'a' versus not 'a'). This is known as the one-versus-all classification. We report an average accuracy of 96% on these analyses and verify that our CNN was indeed able to distinguish between raster plots of different characters.

Our next step was to train a multi-class classifier using a ResNet50 CNN with cross-entropy loss to predict which one of the 31 characters is represented by a given raster plot. Our model achieved 86% accuracy on this task, indicating that the information encoded in the spatial representation of the raster plot was sufficient for a model to learn with reasonable accuracy how to distinguish between all characters.

To identify patterns in the raster data that were discovered by the CNN during training and which may indicate groups of functionally connected neurons, it was necessary to visualize the model's activation map. To do that, we applied a technique known as Gradient-weighted Class Activation Mapping (Grad-CAM) to each trained model [9].

Grad-CAM is a valuable tool for AI explain-ability which produces heat maps indicating which areas in the image passed through a CNN were most important in forming the CNN's prediction for that image. Grad-CAM works by isolating the gradients in the final convolutional layer of a model taken with respect to the class that the model predicted and mapping them to the size and shape of the input image. The full dataset of single-character recordings was passed through the multi-class model and each one-versus-all model, and a heat map was generated using Grad-CAM for each recording.

The heat maps indicate that there exist groups of neurons whose coordinated activity accurately predicts the character classification of a particular raster plot.

3. EXPERIMENTS

For one-versus-all training, we trained 31 models, one for each character, using ResNet50 as the backbone and a fully connected classifier layer with a binary output indicating whether or not the input represented a particular character or one of the other 30. The models used binary cross-entropy loss and were trained using Adam optimizer [22] for 1 epoch. Each model was trained and evaluated with five different starting seeds and five different train-test splits to ensure the classification accuracies they yielded were reproducible. Each train-test split consisted of 117 examples of the given character and 117 examples chosen randomly out of the other characters. These were divided evenly into a training set of 187 examples and a testing set of 47 examples.

The accuracy of the 31 models, averaged over all five experiments is shown in Figure 3, where the standard deviation is indicated in parentheses. For example, the average accuracy of the 'b' versus not-'b' model (evaluated on a dataset containing half examples of the letter 'b' and half examples of random non-'b' characters) over five experiments was 97.44% with a standard deviation of 2.3%. In addition, the average accuracy across all 31 models was 96% with a standard deviation of 2.5%.

For the multi-class problem, a model was trained using ResNet50 as the backbone and a fully connected classifier layer with 31 output classes corresponding to the 31 characters (a through z plus five punctuation marks) written by the subject in Willet et al. The model was trained and evaluated with five different starting seeds and five different train-test splits. The model used cross-entropy loss and was trained using the Adam optimizer for a maximum of 100 epochs using early stopping to avoid over-fitting.

The train-test split for the multi-class experiments consisted of 117 examples of each class, for a total of 3627 examples divided into a training set of 3000 and a test set of 627, evenly distributed across classes.

The model performs with an average classification accuracy over five experiments of 87.08% with a standard deviation of 0.49%. The confusion matrix shown in Figure 1 demonstrates that classification accuracy for each class is reasonably high, falling within a range of 58% for the letter 'n' and 100% for the greater-than character which was used by the subject to denote a space between words. It should be noted that there is very little confusion between classes. The most confused characters are those that are visually similar, such as a 26% misclassification of the character 'n' as 'r', a 25% misclassification of the character 'h' as 'n', and a 19%

Character	Avg (std)	Character	Avg (std)	Character	Avg (std)	Character	Avg (std)
a	0.868 (0.041)	i	0.987 (0.029)	q	0.983 (0.028)	y	0.962 (0.023)
b	0.974 (0.023)	j	0.974 (0.018)	r	0.962 (0.023)	z	0.953 (0.018)
c	0.974 (0.01)	k	0.949 (0.044)	s	0.983 (0.038)	>	0.957 (0.072)
d	0.966 (0.024)	l	0.97 (0.019)	t	0.928 (0.051)	,	0.991 (0.012)
e	0.979 (0)	m	0.97 (0.019)	u	0.974 (0.035)	'	0.906 (0.041)
f	0.991 (0.012)	n	0.961 (0.038)	v	0.987 (0.012)	~	0.953 (0.028)
g	0.957 (0.034)	o	0.979 (0)	w	0.932 (0.035)	?	0.936 (0.015)
h	0.936 (0.015)	p	0.966 (0.012)	x	0.974 (0.023)	**Average**	**0.96 (0.025)**

Table 1. The average accuracy of the binary (one-versus-all) classification for each of the 31 characters where each model is trained and evaluated with five random initializations. Standard deviations are indicated in parentheses. For example, the average accuracy for the 'g' versus not-'g' model was 95.74% with a standard deviation of 1.5%. The average accuracy of all models is 96%.

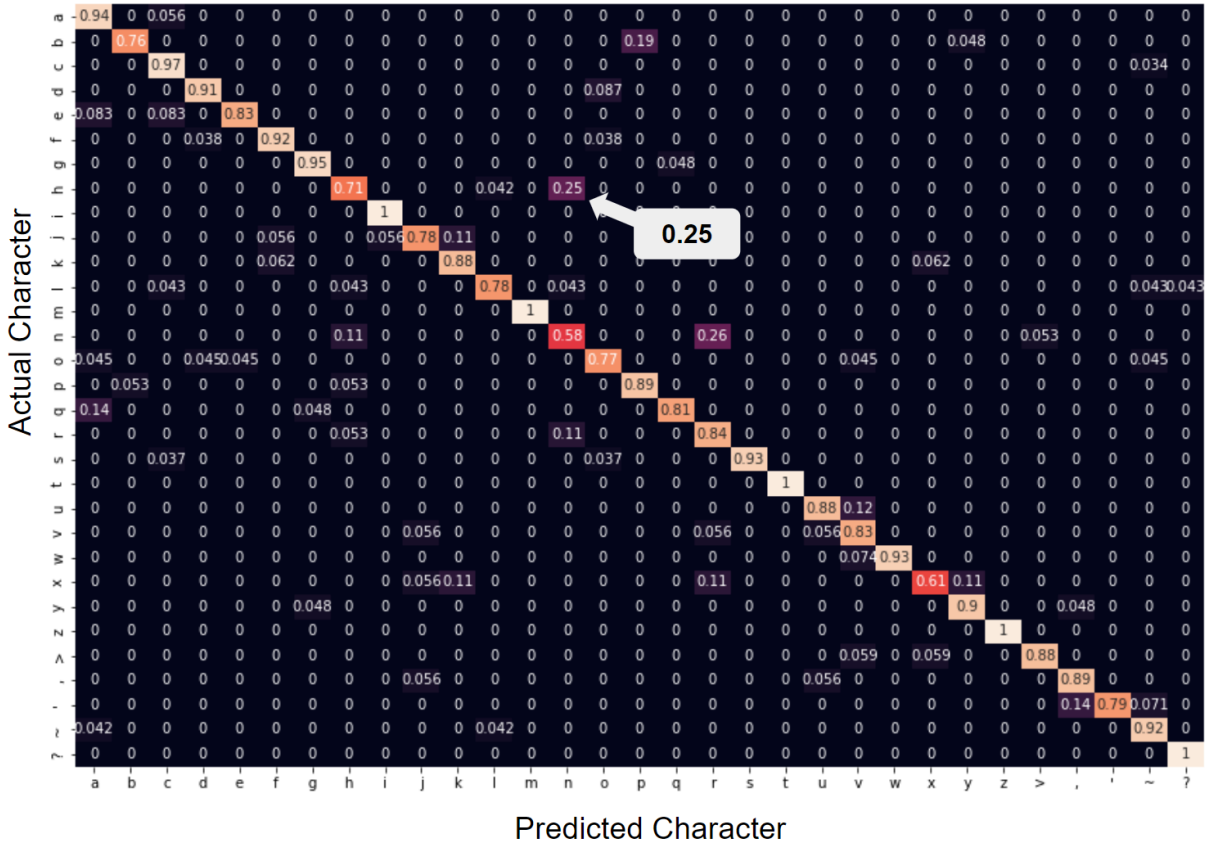

Fig. 1. Confusion matrix for multi-class experiments indicating high classification accuracy for each character, ranging from 58% for the letter 'n' to 100% for 'i', 'm', 't', 'z', and '?'. The off-diagonal non-zero entries represent ill-detected characters. The arrow indicates that the character 'h' is confused for the character 'n' in 25% of the test examples (note, that the imaginary pen strokes for writing 'h' and 'n' characters could be quite similar), these two characters were also represented in interleaved clusters in Willett *et al.* [8] Figure 1. Similarly, 25% of the 'n' characters are confused with 'r'.

misclassification of the character 'b' as 'p'.

We speculate that those most-confused characters are so confused because the mechanics of writing those characters are similar and are thus encoded similarly in the brain. For example, the pen movements in writing the letter 'r' are encompassed in the pen movements in writing the letter 'n', and the letter 'c' is similarly embedded in the letter 'a'.

Experiments using Grad-CAM. As mentioned earlier and as can be seen in 3, we also implemented Grad-CAM by attaching backward hooks to the final convolutional layer of

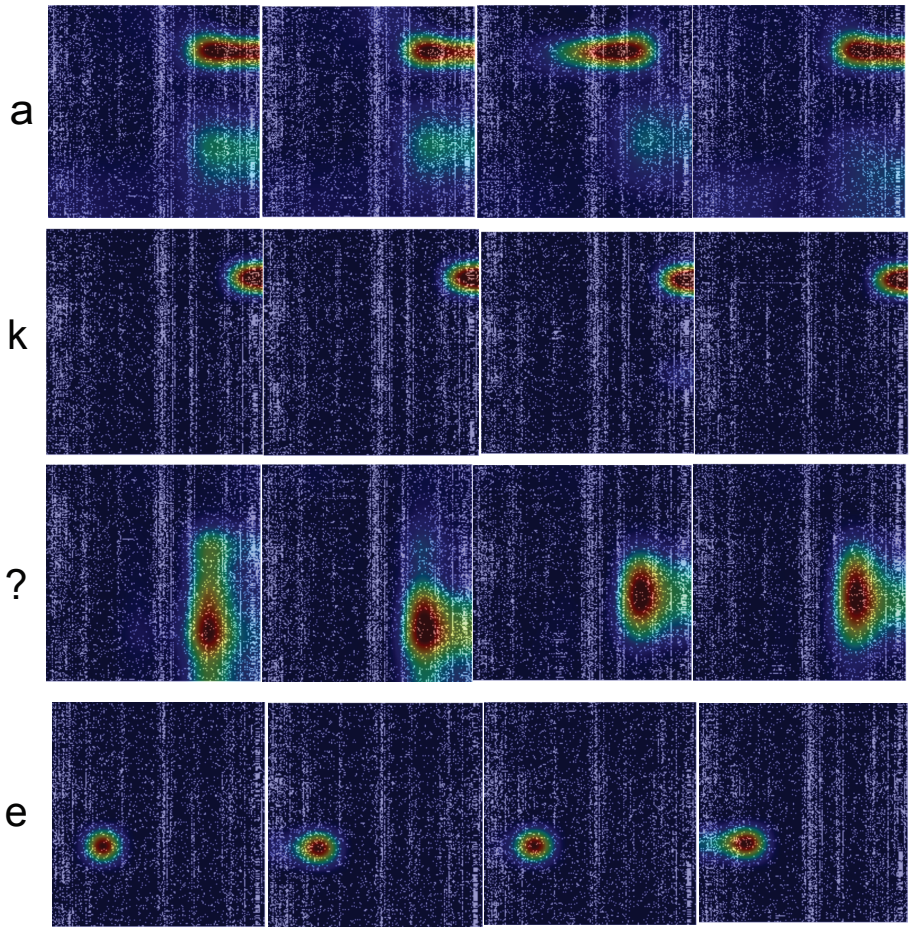

Fig. 2. Grad-CAM images of our one-versus-all binary classification models. We can see the groups of neurons activated together in typical patterns that the CNN recognizes to classify each character.

each of our trained models. Beginning with the one-versus-all models and then the multi-class model, we applied a forward pass of each raster plot in both the training and testing set through the model, collecting the gradient of the activations at the hook with respect to the true label of the input image. The gradients were processed as two-dimensional arrays and displayed as heat maps indicating areas of high and low activation. We resized and smoothed the heat maps to the dimensions of the original input images and superimposed them so that the heat map indicates the area in the raster plot that informed the model's classification. In Figure 3 one can see that correlated activity between groups of neurons changes dynamically based on the character being represented. For example, the characters 'a' and 'k' share a group of neurons whose activation is a strong indicator of them both, however, in the letter 'a' that activation is correlated with another, spatially distant group. Alternatively, some characters have disjoint groups of neuronal activities which represent them, for example, see that the group of neurons representing 'a' and '?' are relatively disjoint.

4. CONCLUSION

In this study, we trained CNN models using two classification methods with the aim of visualizing the emergence of functional neuronal groups when the subject was engaged in imagining writing English characters. The two different classification methods (binary classification of a character versus other characters, as well as multi-class fine-grained classification of all 31 characters) show positive results when classifying the character the subject is imagining writing from the raster images. The high accuracy of the two computational tasks as well as the qualitative Grad-CAM visualizations strongly suggests that the CNN is reliable and accurate in detecting the events where neurons form alliances with other neurons as a principle of computing in the human brain on the time scale of milliseconds. In conclusion, we speculate that these patterns represent the waxing and waning of thoughts in the brain.

5. REFERENCES

[1] Kenneth D Harris and Gordon MG Shepherd, "The neocortical circuit: themes and variations," *Nature neuroscience*, vol. 18, no. 2, pp. 170–181, 2015.

[2] Xiao-Jing Wang and Henry Kennedy, "Brain structure and dynamics across scales: in search of rules," *Current opinion in neurobiology*, vol. 37, pp. 92–98, 2016.

[3] Moshe Abeles, *Corticonics: Neural circuits of the cerebral cortex*, Cambridge University Press, 1991.

[4] E Vaadia, I Haalman, M Abeles, Hagit Bergman, Y Prut, Hi Slovin, and AMHJ Aertsen, "Dynamics of neuronal interactions in monkey cortex in relation to behavioural events," *Nature*, vol. 373, no. 6514, pp. 515–518, 1995.

[5] Wolf Singer, "Neuronal synchrony: a versatile code for the definition of relations?," *Neuron*, vol. 24, no. 1, pp. 49–65, 1999.

[6] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget, "Neural correlations, population coding and computation," *Nature reviews neuroscience*, vol. 7, no. 5, pp. 358–366, 2006.

[7] Neda Shahidi, Ariana R Andrei, Ming Hu, and Valentin Dragoi, "High-order coordination of cortical spiking activity modulates perceptual accuracy," *Nature neuroscience*, vol. 22, no. 7, pp. 1148–1158, 2019.

[8] Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy, "High-performance brain-to-text communication via handwriting," *Nature*, vol. 593, no. 7858, pp. 249–254, 2021.

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[10] Horace B Barlow, "Single units and sensation: a neuron doctrine for perceptual psychology?," *Perception*, vol. 1, no. 4, pp. 371–394, 1972.

[11] Rafael Yuste, "From the neuron doctrine to neural networks," *Nature reviews neuroscience*, vol. 16, no. 8, pp. 487–497, 2015.

[12] Ari S Morcos and Christopher D Harvey, "History-dependent variability in population dynamics during evidence accumulation in cortex," *Nature neuroscience*, vol. 19, no. 12, pp. 1672–1681, 2016.

[13] Shreya Saxena and John P Cunningham, "Towards the neural population doctrine," *Current opinion in neurobiology*, vol. 55, pp. 103–111, 2019.

[14] Mark Stopfer, Vivek Jayaraman, and Gilles Laurent, "Intensity versus identity coding in an olfactory system," *Neuron*, vol. 39, no. 6, pp. 991–1004, 2003.

[15] Patrick T Sadtler, Kristin M Quick, Matthew D Golub, Steven M Chase, Stephen I Ryu, Elizabeth C Tyler-Kabara, Byron M Yu, and Aaron P Batista, "Neural constraints on learning," *Nature*, vol. 512, no. 7515, pp. 423–426, 2014.

[16] Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli, "A theory of multineuronal dimensionality, dynamics and measurement," *BioRxiv*, p. 214262, 2017.

[17] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex," *nature*, vol. 503, no. 7474, pp. 78–84, 2013.

[18] Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen Ryu, Krishna V Shenoy, and Maneesh Sahani, "Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity," *Advances in neural information processing systems*, vol. 21, 2008.

[19] John P Cunningham and Byron M Yu, "Dimensionality reduction for large-scale neural recordings," *Nature neuroscience*, vol. 17, no. 11, pp. 1500–1509, 2014.

[20] Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy, "Neural population dynamics during reaching," *Nature*, vol. 487, no. 7405, pp. 51–56, 2012.

[21] Ben Engelhard, Ran Darshan, Nofar Ozeri-Engelhard, Zvi Israel, Uri Werner-Reiss, David Hansel, Hagai Bergman, and Eilon Vaadia, "Neuronal activity and learning in local cortical networks are modulated by the action-perception state," *bioRxiv*, p. 537613, 2019.

[22] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.