Can recurrent models know more than we do?

Noah Lewis
Georgia Tech
TReNDS
Atlanta, GA, USA

Robyn Miller
TReNDS
Georgia State University
Atlanta, GA, USA

Harshvardhan Gazula
Princeton Neuroscience Institute
Princeton, NJ, USA

Md Mahfuzur Rahman
TReNDS
Georgia State University
Atlanta, GA, USA

Armin Iraji
Georgia State University
TReNDS
Atlanta, GA, USA

Vince. D. Calhoun
TReNDS
Georgia State University
Atlanta, GA, USA

Sergey Plis
TReNDS
Georgia State University
Atlanta, GA, USA

***Abstract*—Model interpretation is an active research area, aiming to unravel the black box of deep learning models. One common approach, saliency, leverages the gradients of the model to produce a per-input map highlighting the features most important for a correct prediction. However, saliency faces challenges in recurrent models due to the "vanishing saliency" problem: gradients decay significantly towards earlier time steps. We alleviate this problem and improve the quality of saliency maps by augmenting recurrent models with an attention mechanism. We validate our methodology on synthetic data and compare these results to previous work. This synthetic experiment quantitatively validates that our methodology effectively captures the underlying signal of the input data. To show that our work is valid in a real-world setting, we apply it to functional magnetic resonance imaging (fMRI) data consisting of individuals with and without a diagnosis of schizophrenia. fMRI is notoriously complicated and a perfect candidate to show that our method works even for complex, high-dimensional data. Specifically, we use our methodology to find the relevant temporal information of the subjects and connect our findings to current and past research.**

***Index Terms*—Machine Learning, Deep Learning, model interpretability, neuroimaging**

I. Introduction

Recurrent neural networks, more specifically long short-term memory (LSTM) models are effective for capturing dynamic information [1], which is a key aspect of medical imaging. However, like all deep networks, they are far too complex and opaque to be readily understood without the post-hoc application of additional model introspection. One popular model introspection method is saliency analysis [2]–[4]. Saliency analysis calculates the change in the prediction of the correct class label with respect to each feature of the input $\frac{\partial y_c}{\partial x}$ using backpropagation for efficiency [5]. These gradients, or "saliency maps" highlight the input features the model output is most sensitive to and what changes will have the largest effect on prediction. Although saliency is one of the oldest methods of neural network model introspection [6], [7], it is an effective and intuitive way to understand neural networks and have spurred recent research, including criticisms and validation studies [8], [9]. Importantly, unlike the alternatives, gradient-based saliency passes all validity checks required for a good model introspection technique [8]. Given effective ways of interpreting complex over-parameterized models, understanding the nuances of why a model makes certain decisions about the input data may improve our understanding of these data. This perspective may illuminate subtleties of medical imaging data that linear models miss, and gives researchers a broader insight into complex systems such as the human brain. However, saliency is not a practical tool for explaining dynamic models such as LSTM networks [10] because of a phenomenon known as "vanishing saliency" [11], [12], expressed as a drastic reduction of gradient magnitudes towards the earlier time steps. We propose a way to bypass this problem and leverage the power of saliency to properly examine LSTM's decisions. We show that attention, a powerful technique [13] for sequence representation, can effectively mitigate this problem in a minimally invasive way for the original model. To the best of our knowledge, this approach has not been used before to enhance LSTM introspection.

Like most model introspection techniques, our approach is sensitive to random initialization [14] and we take advantage of that sensitivity. To create robust and effective saliency maps, we bootstrap a distribution of trained parameters, and then select the resulting saliency maps that are closest to the average of this distribution. In other words, we use the same architecture with 300 random initializations to create a bootstrapped distribution of the model. We compare our approach to the only other research that we know studies the problem of "vanishing saliency", input-cell attention [10], and show that our approach more accurately and effectively captures the truly relevant information within the input data.

To validate our methodology, we perform an experiment using a synthetic dataset, uniquely designed to reveal and quantify the class discriminatory information. We also analyze functional magnetic resonance imaging (fMRI) data from individuals with and without a diagnosis of schizophrenia [15]. Specifically, we use independent component analysis (ICA) to reduce the feature size of the fMRI data (which can include 10s of thousands of features per time step), input the ICA timecourses into our model, and then compute the saliency maps. These saliency maps are then analyzed to uncover group

differences that might otherwise be hidden. We study the group differences based on the most salient temporal information. In other words, we use our maps to find the most salient time steps, and group these into blocks or windows of salient steps. Finally, we find the group differences based on these blocks, and then analyze the relationship between these block patterns and the diagnostic scoring system known as the Positive and Negative Syndrome Scale (PANSS) for schizophrenia [16]. PANSS, a quantifying metric for the disorder that includes many distinct symptoms, each with a severity score. We use our salient block information to find specific symptoms more closely related to our salient information. These experiments show that our method can be used on complicated and nuanced real world data with scientifically relevant or even novel results.

II. Methods

A. Our Approach

To compensate for vanishing saliency, we insert an additive attention mechanism [13] between the outputs of a bi-directional LSTM (bi-LSTM) [17] and the final output layer, which creates a direct gradient flow path from the classification to the input via the attention parameters. A bi-directional LSTM was chosen because we do not consider streaming data, the additional parameters aid training, and the extra directional flow for gradients may also improve the quality of the maps.

The attention mechanism [13] is a powerful way to amalgamate temporal information and "attend" only to the most important steps in the LSTM output by assigning a weight to each step. To parameterize the attention mechanism, at each step, we pass the LSTM output through an attention network of two feed-forward layers to create a single, per-step weight value. The weight values from all time steps are jointly softmaxed and used to adjust the LSTM output at the individual time steps. As the model is bi-directional, we use the output from both the forward and backward directions concatenated into a single vector as our context for the attention mechanism. In other words, $h\_backward_T$ is concatenated with $h\_forward_T$ and passed through the attention mechanism to give us the respective attention weight for that time step. After weighting the hidden outputs by the attention scalar, they are summed along the time dimension and pushed through a linear transform for classification.

The saliency maps for each sample are calculated from the trained models, and indicate which features are most relevant to the model's accurate predictions. However, we observe that the maps are rarely stable, and vary widely with the initial randomization. To correct this, we train multiple models (keeping the same hyperparameters) with different random initializations and calculate saliency maps from each model (for all experiments, we train 300 total models). For each input sample, we select the map that is closest to the average map over all models for that sample, using Euclidean distance. Figure 1 is a diagram of our procedure.

B. Input-Cell Attention

Input-cell attention [10] is the only other prior work for mitigating vanishing saliency known to us, which leverages an attention mechanism to weight the input before it is processed by the LSTM. Each input step into the LSTM has its own attention parameterization mechanism (i.e. weight matrices), in which each time step (x_t) is weighted by the matrix A_t, where $A_t = softmax(W_2 tanh(W_1 X_t^T))$ and W_1 and W_2 are learned matrices, giving us the input to the LSTM, $M_t = A_t X_t$ for each step, *t*. This method uses an input matrix X_t of size N x t, where N is the number of features per time step, and t is the number of time steps. The first trained weight matrix, W_1 is of size N x d_a, where *N* is the number of features per time step and d_a is a hyper parameter. W_2 is a r x d_a matrix, where r is a window of input time steps that the model attends to, making A_t (r x N size) the final input to the LSTM. There are two issues with this method. Firstly, it requires many more parameters than an LSTM with attention on the output, where a static number of new parameters are required for each step within a window (r steps), making it slower and more memory intensive. Secondly, we show that our work is quantitatively more effective at capturing the truly relevant features, using the same metrics as in [10].

C. Synthetic Data

The synthetic data is specifically engineered so that the relevant information within the data is quantifiable and interpretable. In this dataset of 30,000 samples, each sample is randomly generated as Gaussian noise (μ=0, σ=1) with a sequence length of 200 and 30 features, then randomly assigned a class label of either 0 or 1. As we want to show that our method can capture complex dynamics, we devise a method to perturb the inputs with hidden class-relevant dynamic information. Vector autoregression (VAR) is used to control the underlying dynamics of each sample. Rank p VAR model explains the evolution of a variable over time with the generalized equation: $x_t = c + A_1 x_{t-1} + A_2 x_{t-2} + ... + A_p x_{y-p} + e_t$. In our experiments we set $p = 1$, which resulted in a single $\mathbf{A}$ matrix. For all samples, the VAR is computed using a positive semi-definite matrix, $\mathbf{A}$. Then, 15 successive steps are randomly chosen to be perturbed with new dynamic information. Or, two more positive semi-definite matrices, B and C, are created and VAR is again used to compute 15 new steps using Gaussian noise and either matrix B or C, depending on the class label of the sample. These new steps, $x'_{t:t+15}$ are added to the sample at a randomly selected interval ($x_{t:(t+15)}$), with an interpolation variable, α resulting in the equation: $\alpha x'_{t:(t+15)} + (1-\alpha) x_{t:(t+15)}$.

D. Saliency Accuracy Measures

Since the relevant information of the synthetic datasets is easily quantifiable, we can use basic similarity scores between the saliency maps and proper representations of the input to understand the quality of the maps. To be fair to both our method and input-cell attention, we only compare the samples that were held out during model training. Firstly, as the input

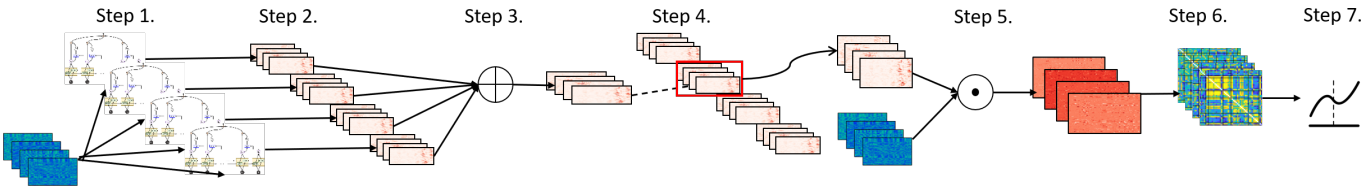

Fig. 1. Flowchart describing our pipeline for analyzing the ICA timecourses. For all other datasets, we use only the first 4 steps to calculate the finalized maps. Step 1: we train 300 separate models (each with the same architecture) using different random initializations for each model on the same set of ICA timecourses. Step 2: we calculate the saliency maps for each sample from all 300 models. Step 3: We calculate the average saliency map for each sample over all 300 models. Step 4: We select the per-model set of maps with the lowest Euclidean distance to the average over all models. Finally resulting in a stable saliency map for each input sample.

data is noisy, we need a reasonable representation of each sample. For the VAR experiment, we represent each sample as a binary matrix in which only the elements within the perturbed regions are ones, and all other elements are zero. The 15x30 window perturbed with added dynamics is set to all ones, leaving all other features as zero. Additionally, for the saliency maps from both our method and input-cell attention, we pass each sample through an absolute function [18]. To conduct a fair comparison with [10], we use both of the similarity metrics therein: Euclidean distance and weighted Jaccard. We also evaluate the sum of all salient values within the window over the sum of the entire map.

To ensure an unbiased sampling of the timecourses with our model, we separate the data into 27000 training samples and 3000 test samples. We train 300 models on the non-holdout set and generate the maps for every sample, then select the saliency maps using the selection criteria described in our approach section and generate the saliency maps for the holdout set as well. We chose 300 due to computational restrictions, as each model can take some time to train. These maps are then fed through either a rectified linear unit (ReLU) function or absolute function (depending on the experiment) to avoid relying on both positive and negative derivatives for the relevant information. Recent research has shown that removing negatives entirely from saliency can be beneficial [19].

E. Salient fMRI Time Courses

The fMRI data consists of 313 subjects, those with and without a diagnosis of schizophrenia from The Function Biomedical Informatics Research Network (FBIRN) data repository [20], [21]. The data is encoded using ICA to create 47 timecourses for each subject [22], [23]. Our goal is to find patterns within these timecourses as identified by the saliency maps. For this experiment, our architecture is a bi-LSTM with 50 hidden units in each direction (100 total), and two linear layers to parameterize the attention weights (the first linear weight has 50 hidden units, the second weight has 1 output node), and a learning rate of .001. As this dataset is much smaller than our synthetic data, we use 10-fold cross validation. Given our methodology, we randomly generate 300 models for each fold. Then, we select the models from the fold which have the highest average holdout accuracy. Finally, we use our methodology on the saliency maps from these 300 randomly generated models to select the best per-subject map.

F. Finding Group Differences Between Patients and Controls

As an initial investigation of the saliency maps from ICA timecourses, we evaluate the temporal aspects of the maps. Because the maps have visually apparent blocks of salient steps, we convert the maps into 1-dimensional vectors by summing along the component axis, defining each time step as either salient or non-salient. For each subject's map, we absolute the maps and sum them along the component axis. We binarize each step as salient if it is above the subject-wise mean and non-salient otherwise. Following this, we group the time steps into blocks of either salient or non-salient. We consider a block to be any number of sequential steps. Since our goal is to better understand the differences between patients and controls, we analyze the lengths of these blocks. We suggest that each block has some pattern that is learned by the model, and that the block length is the temporal aspect of this pattern. In order to quantify these block lengths, we find the median block length per subject, and then statistically compare these median block sizes between the two groups.

G. PANSS Scores and Symptom Analysis

PANSS is a widely-used scale for measuring the symptom intensity of patients with schizophrenia. As the name, Positive and Negative Syndrome Scale suggests, it covers both positive and negative syndromes of the disorder. Positive symptoms are those symptoms that increase the severity or warp typical functioning, such as hallucinations, grandiose thinking, and hostility. Negative symptoms are those in which the typical functioning is diminished, such as a reduction in abstract thinking, poor rapport with others, and blunted affect. Along with the schizophrenia specific symptoms, there are also 16 general psychological symptom scores, include anxiety, depression, and motor retardation.

Our goal is to find the relationships between the salient block sizes and both the positive and negative scores, excluding the general psychological scale. These positive symptoms are: conceptual disorganization, delusions, manic-like excitement, grandiose thinking, hallucinatory behavior, hostility, and suspiciousness. In order to analyze these relationships, we consider the median block size of each subject and compare to all available scores. This entails using multiple regression for all 7 positive scores and all 7 negative scores as our independent variables, and the median block size as our dependent variable. Along with these 14 symptom scores, we also

regress out four confounding variables: the average general psychological symptom scores, age, gender, and head motion. Head motion is a particularly important confounding variable as it contributes noise to the MRI scan, and can, according to previous research, correlate with a diagnosis of schizophrenia [24]. After calculating the regression coefficients, we analyze the statistical significance of the effect from each variable to pinpoint and examine symptoms related to the block sizes. In our case, 148 patients with schizophrenia had PANSS scores, and there were no scores for typical controls. After computing the multiple linear regression, we found the positive scores that were most significantly effected by the median block size.

After finding any scores with a significant relationship to the median block size, we fit our findings into current research and evaluate the novelty of our findings.

TABLE I

Comparing the LSTM+attention with input-cell attention on the VAR dataset. WITHIN EACH CELL IS THE AVERAGE AND STANDARD DEVIATION OF THE METRIC OVER 3,000 TEST SAMPLES, AND THE p-VALUE COMPARING THE TWO MODELS USING A 2-SAMPLE t-TEST.

	VAR Dataset		
	Euclidean Distance	Overlapping Values	Weighted Jaccard
LSTM + Attention	μ=4.57, σ=1.16, $p<$ .0001	μ=.56, σ=.31, $p<$ .0001	μ=.10, σ=.05, $p<$ .0001
Input-Cell Attention	μ=2.35, σ=.45, $p<$ .0001	μ=.15, σ=.05, $p<$ .0001	μ=.07, σ=.02, $p<$ .0001

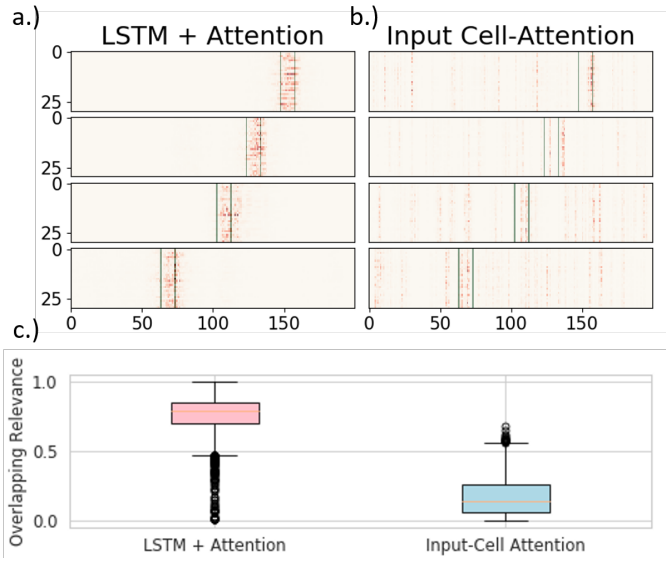

Fig. 2. The results from the analysis of the VAR dataset. Several examples of input samples matched with the saliency maps from the LSTM+attention (a) and input-cell attention (b). c: The boxplots of the overlapping values over all 3,000 holdout samples for both models. The overlap is defined as the percentage of the total sum of the maps that is within the relevant area over the total sum of the map. The green blocks in a and b highlight these relevant regions. Each baseline image has a certain underlying transition matrix, as computed by VAR, and each sample is interpolated with one of two different transition matrices, depending on the class label (the label along the y axis) within the highlighted relevant area. Notice the difficulty in which it can be to visually determine where the relevant information occurs.

III. RESULTS

A. Synthetic Data

To ensure stability of the maps, we chose neural network hyperparameters that lead to the highest accuracy for both the LSTM+attention and cell-attention models. We found, after a non-exhaustive search, that for both models, 50 hidden units for the LSTM was the most appropriate number of hidden units. We also found that the attention mechanism in the LSTM+attention did well with two layers, with 50 hidden units and 1 output unit, respectively. The input-cell attention achieved a test accuracy of 0.90 on the VAR dataset. Over all initializations of the LSTM+attention, the average accuracy on the VAR test data was 96%. The statistical comparison measurements of both our method and input-cell attention can be found in table I. Several randomly selected examples of the maps for both methods as well as a break down of the saliency accuracy are in figure 2.

B. Group Difference Results

Visual inspection of the saliency maps show some temporal "spotlights", or temporal blocks of highly salient information followed by blocks with little or no relevant information. We then threshold each subject's map, with the subject-wise mean, to give a series of salient and non-salient temporal blocks, identified as a binary vector (following a component-wise absolute summation as described earlier). Each block of continually salient steps (in the binary vector) is considered to be one "block". We find that contiguous temporal blocks of median saliency levels are significantly shorter in patients: patient average block length = 14.9 timepoints, while in controls the average block length was 19.5 steps long. A two-sample t-test showed that this difference is highly significant ($p<0.0004$). We could speculate that this may relate to previous findings that patients exhibit shorter transient periods of disrupted activity [25]–[27].

C. PANSS Analysis Results

Comparing the average block size of the 148 subjects with available PANSS scores, we found a statistically significant relationship between block size and one symptom, excitement. Excitement is a positive symptom, and characterizes poor impulse control, hyperactivity, hostility, and uncooperativeness. With a p-value of .0297, our regression coefficient is -3.15. The negative coefficient shows that smaller blocks are more related to high excitement scores, indicating that the role of shorter-duration sequential patterns in correctly classifying the disorder is amplified in patients with high "excitement" scores.

IV. DISCUSSION

Saliency can be noisy and sometimes unstable. We show empirically that by building a distribution of possible initializations, training each randomly initialized model, and then selecting the saliency map (for each sample) from a given model that is closest to the average maps, this issue can be rectified. Our VAR dataset shows that, even with complex and dynamic data, our method produces saliency maps that

quantifiably find the truly relevant information within the dataset. Thorough analysis of the ICA saliency maps reveal several interesting insights. When grouping the maps into blocks of salient and non-salient information, we find that the control group had significantly fewer blocks. This corresponds to our finding that the controls also had significantly longer blocks of salient information. This may have a wide array of implications, however, we will point the reader towards research on reduced connectivity in the brain of patients with schizophrenia [28]–[30]. Essentially, it is possible that our results relate to current research showing that fMRI data of patients with schizophrenia is less cogent and stable.

Finally, we hypothesized that these blocks are somehow related to specific symptoms of the disorder, and not just the disorder itself. We found a clear relationship between the median per-subject block sizes and the excitement symptom. Research shows that excitement can be observed in fMRI scans in both task-based experiments [31] and resting-state experiments [32]. One important caveat is that as we do not have an associated sub-scoring system, our results may be associated with more specific aspects of excitement. In summary, we highlight the significant challenges posed by saliency maps owing to the vanishing saliency phenomenon and propose the addition of attention mechanisms to mitigate this drawback. Satisfactory results on both synthetic and real data highlight the utility of this work in better understanding the temporal patterns of fMRI. This work could easily be expanded to different disorders and could have significant implications in a clinical setting for diagnostic purposes.

REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] N. J. S. Morch *et al.*, "Visualization of neural networks using saliency maps," in *Proceedings of ICNN*, vol. 4, 1995, pp. 2085–2090 vol.4.

[3] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep cnn features," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5012–5024, 2016.

[4] Y. Wang, H. Su, B. Zhang, and X. Hu, "Learning reliable visual saliency for model explanations," *IEEE Transactions on Multimedia*, vol. 22, pp. 1796–1807, 2020.

[5] R. Nath, B. Rajagopalan, and R. Ryker, "Determining the saliency of input variables in neural network classifiers," *Comps. & Ops. Research*, vol. 24, no. 8, pp. 767 – 773, 1997.

[6] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th Intl. Conf. on data science and advanced analytics*. IEEE, 2018, pp. 80–89.

[7] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 10 2017.

[8] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, 2018, pp. 9505–9515.

[9] P.-J. Kindermans *et al.*, "The (un) reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 267–280.

[10] A. A. Ismail, M. K. Gunady, L. Pessoa, H. C. Bravo, and S. Feizi, "Input-cell attention reduces vanishing saliency of recurrent neural networks," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, 2019, pp. 10 813–10 823.

[11] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[12] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[14] A. Atakulreka and D. Sutivong, "Avoiding local minima in feedforward neural networks by simultaneous learning," in *AI 2007: Advances in Artificial Intelligence*, M. A. Orgun and J. Thornton, Eds., 2007, pp. 100–109.

[15] M. S. Salman *et al.*, "Group ica for identifying biomarkers in Schizophrenia: 'adaptive' networks via spatially constrained ica show more sensitivity to group differences than spatio-temporal regression," *NeuroImage: Clinical*, vol. 22, p. 101747, 2019.

[16] S. R. Kay, A. Fiszbein, and L. A. Opler, "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia," *Schizophrenia Bulletin*, vol. 13, no. 2, pp. 261–276, 01 1987.

[17] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Sig. Proc.*, vol. 45, pp. 2673 – 2681, 12 1997.

[18] N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos, "On computational modeling of visual saliency: Examining what's right, and what's left," *Vision Research*, vol. 116, pp. 95 – 112, 2015, computational Models of Visual Attention.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, 2019.

[20] D. B. Keator, T. G. van Erp, J. A. Turner, and et al., "The function biomedical informatics research network data repository," *NeuroImage*, vol. 124, pp. 1074 – 1079, 2016.

[21] Z. Fu *et al.*, "Characterizing dynamic amplitude of low-frequency fluctuation and its relationship with dynamic functional connectivity: an application to schizophrenia," *Neuroimage*, vol. 180, pp. 619–631, 2018.

[22] E. Damaraju *et al.*, "Dynamic functional connectivity analysis reveals transient states of dysconnectivity in Schizophrenia," *NeuroImage: Clinical*, vol. 5, pp. 298 – 308, 2014.

[23] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar, "A method for making group inferences from functional mri data using independent component analysis," *Human brain mapping*, vol. 14, no. 3, pp. 140–151, 2001.

[24] C. Makowski, M. Lepage, and A. Evans, "Head motion: the dirty little secret of neuroimaging in psychiatry," *Journal of psychiatry & neuroscience: JPN*, vol. 44, pp. 62–68, 01 2019.

[25] J. A. Turner *et al.*, "Reliability of the amplitude of low-frequency fluctuations in resting state fmri in chronic schizophrenia," *Psychiatry Research: Neuroimaging*, vol. 201, no. 3, pp. 253 – 255, 2012.

[26] J. Turner *et al.*, "A multi-site resting state fmri study on the amplitude of low frequency fluctuations in schizophrenia," *Frontiers in Neuroscience*, vol. 7, p. 137, 2013.

[27] V. D. Calhoun, K. A. Kiehl, and G. D. Pearlson, "Modulation of temporally coherent brain networks estimated using ica at rest and during cognitive tasks," *Human Brain Mapping*, vol. 29, no. 7, pp. 828–838, 2008.

[28] P. Skudlarski *et al.*, "Brain connectivity is not only lower but different in schizophrenia: A combined anatomical and functional approach," *Biological Psychiatry*, vol. 68, no. 1, pp. 61 – 69, 2010.

[29] G. Gifford *et al.*, "Resting state fmri based multilayer network configuration in patients with schizophrenia," *NeuroImage: Clinical*, vol. 25, p. 102169, 2020.

[30] R. L. Miller *et al.*, "Higher dimensional meta-state analysis reveals reduced resting fmri connectivity dynamism in schizophrenia patients," *PLOS ONE*, vol. 11, no. 3, pp. 1–24, 03 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0149849

[31] Y. Nishimura, R. Takizawa, M. Muroi, K. Marumo, M. Kinou, and K. Kasai, "Prefrontal cortex activity during response inhibition associated with excitement symptoms in schizophrenia," *Brain Research*, vol. 1370, pp. 194 – 203, 2011.

[32] Z. He *et al.*, "Aberrant intrinsic brain activity and cognitive deficit in first-episode treatment-naive patients with schizophrenia," *Psychological medicine*, vol. 43, pp. 1–12, 08 2012.