Comparing Traditional Method with Deep Learning in fMRI Pattern Recognition

Adam Huang

Department of Computer Science, Carleton College

(Dated: December 13, 2019)

Deep Learning achieves huge success in various image processing tasks. It is, therefore, a popular candidate for processing fMRI data which encodes rich information of the neural activity and is hard to be interpreted directly with human eyes. We explore the performance of deep learning models compared with traditional machine learning methods on 125 fMRI samples from the ADHD-200 dataset. Concretely, we compare the results of pixel-wise KL Divergence and Deep Learning models in pattern recognition and localization of the fMRI data. We find that the pixel-wise KL Divergence based model consistently outperforms deep learning models in all tasks. We expect that a larger amount of data and more sophisticated data pre-processing would help deep learning models in the future.

I. ADHD-200: A COMPREHENSIVE FMRI SURVEY

The dataset we will work on is called ADHD-200, which is published in 2011 for the 'ADHD-200 Global Competition', which aims at identifying biomarkers of attention-deficit/hyperactivity disorder (ADHD) from resting-state functional magnetic resonance imaging (rs-fMRI) and structural MRI (s-MRI) data [1]. The entire dataset is composed of 973 individuals, where the fMRI part of the raw data was preprocessed by [1].

Traditional machine learning and heuristic-based analysis have been employed to recognize neural activity patterns as biomarkers for ADHD [2] [3], which receives success in finding correlated activity patterns with ADHD. Deep learning models are also employed for classifying the type of ADHD through pattern recognition, including using deep belief net [4], and using CNN [5]. However, these models still perform relatively poor compared to the state-of-the-art model in other image processing tasks.

Our goal here is to compare deep learning with traditional machine learning and statistical analysis on pattern recognition and localization. We first analyze the phenotypic characteristics of ADHD-200 and visualize the fMRI data. We then train a model based on pixel-wise KL Divergence and two deep learning models. Finally, we discuss and compare the models' performances.

II. PRELIMINARY ANALYSIS OF ADHD-200

We analyze phenotypic information and provide visualization for whole the ADHD-200 dataset as well as our selected subset of 125 samples used for our modeling experiment.

A. Phenotypic Characteristics Overview

There are 8 types of phenotypic information recorded for the participant kids: "site", "gender", "handedness", "diagnosis", "ADHD measure", "medical status", "IQ performance", and "quality control". While "site", "quality control", and "ADHD measure" are used for initial data collection, "gender", "handedness", and "IQ performance" are intrinsic characteristics of the kids. "Diagnosis" is an indicator of whether the given kids have ADHD, and "medical status" indicates whether the kids receive any medical treatments before the experiments. We focus on analyzing the general distributions of these phenotypic characteristics, and the correlation, if there are any, between the intrinsic characteristics and the diagnosis results of the kids.

The top figure in Fig. 1 a) shows the age distribution of the participant kids of the entire ADHD-200 dataset. As shown, most of the participant kids are of age from 8 to 14. From the phenotypic dataset, we also know that the gender split of the participant kids is around a half-half split, while more than 80 % of the kids are right-handed.

The middle figure in Fig. 1 a) shows the distribution of Diagnosis Measurement of all the participant kids. A diagnosis of 0 means the kids are typical without ADHD-200. A diagnosis of 2 means the kids have impulsive ADHD, a diagnosis of 3 means they have attentative ADHD, and 1 means the kids have both types of ADHD. As shown, most of the kids in the dataset are normal kids, and kids having only impulsive ADHD is significantly less than kids having the other two types of ADHD.

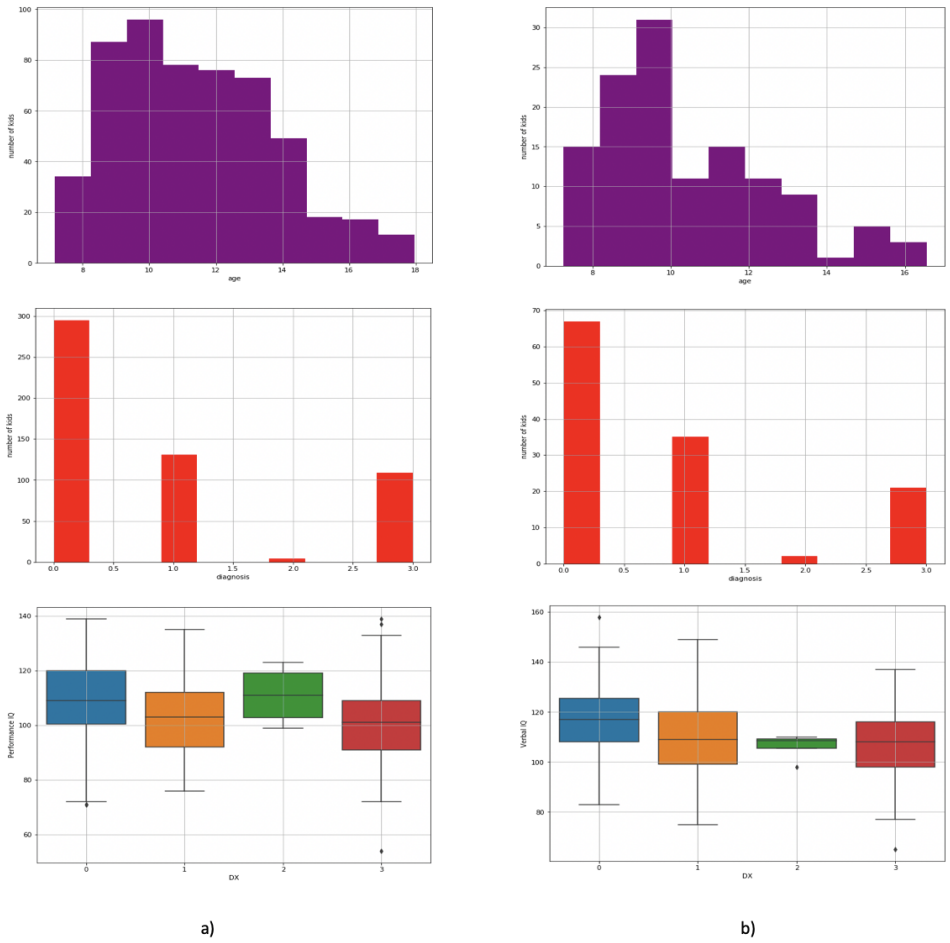

FIG. 1. Phenotypic characteristics of a) the entire ADHD-200 dataset, and b) the selected 125 samples for modeling experiments. The upper plots show the age distributions of the participant kids. The middle plots show the diagnosis types of ADHD of the participant kids. The lower plots show the boxplots of IQ Performance over types of ADHD of the participant kids.

The bottom figure Fig. 1 a) shows the boxplot of IQ Performance versus the type of diagnosis of all the participant kids. Comparing the blue box and the orange, the distribution of IQ Performance of kids having ADHD is lower than the distribution that of the typical kids, indicating that there could be a correlation between ADHD and Performance in IQ tests of the kids.

The phenotypic information of 125 selected samples for our experiments is shown in Fig. 1 b). Each figure is a correspondent of the figure to its right that describes the distributions of the entire ADHD-200 dataset. Comparing the figures in Fig. 1 b) with those in Fig. 1 a), it is shown that our 125 selected samples resemble the distributions of important phenotypic

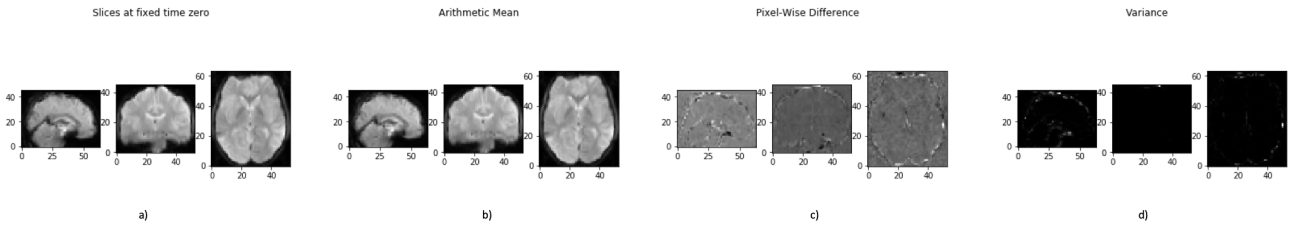

FIG. 2. Visualization of the fMRI data. a) Selected slices of fMRI data at fixed time 0. b) Selected slices of the *Arithmetic Mean* of fMRI data. c) Selected slices of the *Pixel-wise Difference from Arithmetic Mean* of fMRI data. c) Selected slices of the *Pixel-wise Difference from Arithmetic Mean* of fMRI data. c) Selected slices of the *Variance* of fMRI data.

characteristics of the entire dataset.

B. Data Visualization

We visualize the fMRI data of the 125 selected with the following metrics.

1. Arithmetic Mean

Given a time series of fMRI data of shape (x,y,z,t), the arithmetic mean ($\overline{x}$,$\overline{y}$,$\overline{z}$) is

$$(\overline{x}, \overline{y}, \overline{z}) = \frac{1}{\Delta t} \sum_{t=0}^{\Delta t} (x_t, y_t, z_t), \tag{1}$$

where Δt is the lifetime of a time series, and (x_t, y_t, z_t) is the 3D point cloud at given time t.

2. Pixel-wise Difference from Arithmetic Mean

Given a 3D point cloud (x_T,y_T,z_T) at fixed time T of the fMRI time series, the pixel-wise Difference from Arithmetic Mean (x_-,y_-,z_-) is:

$$(x_-, y_-, z_-) = (x_T, y_T, z_T) - (\overline{x}, \overline{y}, \overline{z}). \tag{2}$$

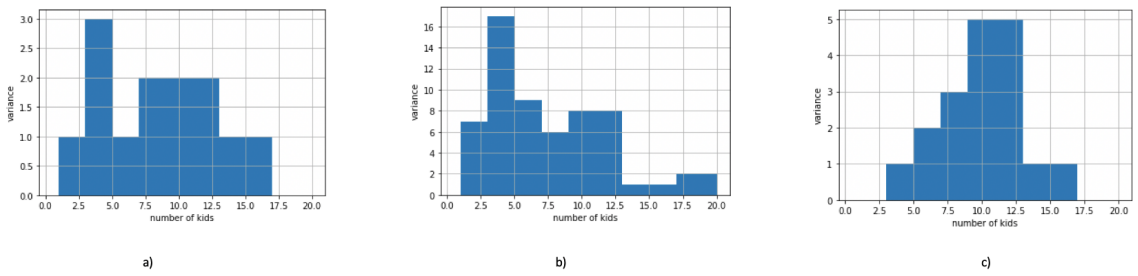

FIG. 3. An illustration of the distribution of variance for each pixel.

3. Variance

Given a time series of fMRI data of shape (x,y,z,t), the variance (x_{σ^2},y_{σ^2},z_{σ^2}) is:

$$(x_{\sigma^2}, y_{\sigma^2}, z_{\sigma^2}) = \frac{1}{\Delta t}\sum_{t=0}^{\Delta t}((x_t, y_t, z_t) - (\overline{x}, \overline{y}, \overline{z}))^2, \tag{3}$$

where Δt is the lifetime of a time series and (x_t, y_t, z_t) is the 3D point cloud at given time t.

The visualization of the fMRI time series for each of the above metrics and the time series at a fixed time stamp along each axis x, y and z, is shown in Fig. 2.

III. MEASURING KL DIVERGENCES IN ADHD-200

We compute pixel-wise KL Divergence of variance between different populations of diagnosis types to identify important neural activity patterns and locations that could be potential indicators for ADHD. We first introduce the exact KL Divergence metric and then discuss the computational results.

A. Pixel-wise KL Divergence of variance

Given two distributions of variances X = (n_x, $(x_{\sigma^2}, y_{\sigma^2}, z_{\sigma^2})_{n_x}$) and Y = ($n_y$, $(x_{\sigma^2}, y_{\sigma^2}, z_{\sigma^2})_{n_y}$), for each n_x and n_y be the individual samples from the two distribution, we downsample both distrbutions into bins of values of variances [0.5,1,3,5,7,9,13,17,20] for each pixel in the variance matrices $(x_{\sigma^2}, y_{\sigma^2}, z_{\sigma^2})_{n_x}$) and $(x_{\sigma^2}, y_{\sigma^2}, z_{\sigma^2})_{n_y}$), and count the number of individual sample in the bins. An illustration in the form of histograms is provided in Fig. 3. Denote the two downsampled pixel-wise distributions as X' and Y' and assume a uniform prior of

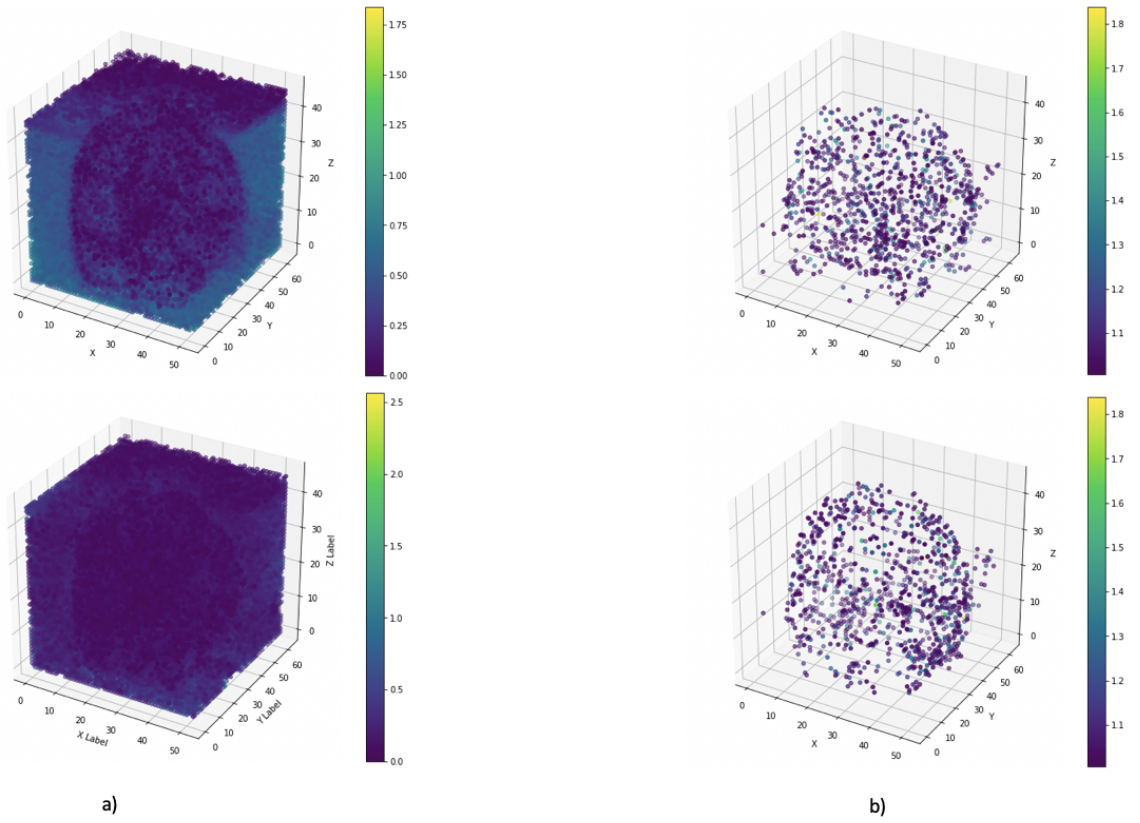

FIG. 4. The pattern recognition and localization results of the pixel-wise KL Divergence method. a) The pixel-wise KL Divergence values of all pixels. b) The pixel-wise KL Divergence values of top 10000 pixels with pixel-wise KL Divergence greater than 1

each individual sample, the pixel-wise KL Divergence is then:

$$D_{kl}(X'||Y') = \frac{1}{n_x} \sum_{b \in bins} C_b(X') log(\frac{C_b(X')n_y}{C_b(Y')n_x}), \tag{4}$$

for $C_b(X)$ being the function that count the number of samples in each bin. A pixel-wise KL-Divergence of values greater than one means the pattern is more likely to correlate with X' than to co-correlate with both X' and Y'.

B. Discussion

We compute pixel-wise KL Divergence for variances between the population of typical kids and populations of kids with different types of ADHD. Since kids with type 2 ADHD has few amounts of data, we only compare pixel-wise variance distribution of type 1, denoted

as X_1, and that of type 3 ADHD, denoted as X_3, with that of typical kids, denoted as X_0. The results are shown in Fig.

In Fig. 4 a), the general distribution of pixel-wise KL Divergence across the entire 3D point cloud is shown. One obvious observation comparing the upper and lower image is that the pixel-wise KL divergence of outer-shell pixels between type 1 and typical kids $D_{kl}(X_1 \,||\, X_0)$ is much higher than that between type 3 and typical kids $D_{kl}(X_3 \,||\, X_0)$. However, this difference is likely because of background noises, as the pixel-wise KL Divergence value of the outer-shell pixels are in general small (less than 1). Across the entire 3D point cloud, most of the pixel-wise KL divergence values are less than 1, meaning that most of the neural activity patterns are likely the same for both typical and ADHD-carrying kids.

Apart from visualizing the general distribution, we also look at the top 10000 pixels with pixel-wise KL Divergence. Interestingly, most of the pixels with high KL Divergence, for both $D_{kl}(X_1 \,||\, X_0)$ and $D_{kl}(X_3 \,||\, X_0)$, are presented around the inner center, as illustrated in Fig. b). Such a pixel pattern, as we propose, is very likely a neural activity pattern because of its local concentration. Although most of the pixel-wise KL Divergences for these pixels are barely above 1, they present in the same regions as where the pixels with absolutely high confidence (pixel-wise KL Divergence greater than 1.5) to be only correlated with type 1 are presented. We encourage a statistical significance test of our claim.

IV. DEEP LEARNING APPROACH TO ADHD-200

Our task is to classify the diagnosis type of the kids given their fMRI data. We present two deep learning models of slightly different architectures and train the two models with data obtained from two different metrics. For pattern localization, we will also look at the feature maps of each convolutional layers.

A. Architectures

A visualization of the two architectures can be found in Fig. 6.

The first model splits the 3D point cloud into three different feature extractors. Each feature extractor takes as input an averaged image of the variance matrix along a specific axis. The outputs of the feature extractors are then flattened and concatenated into a single

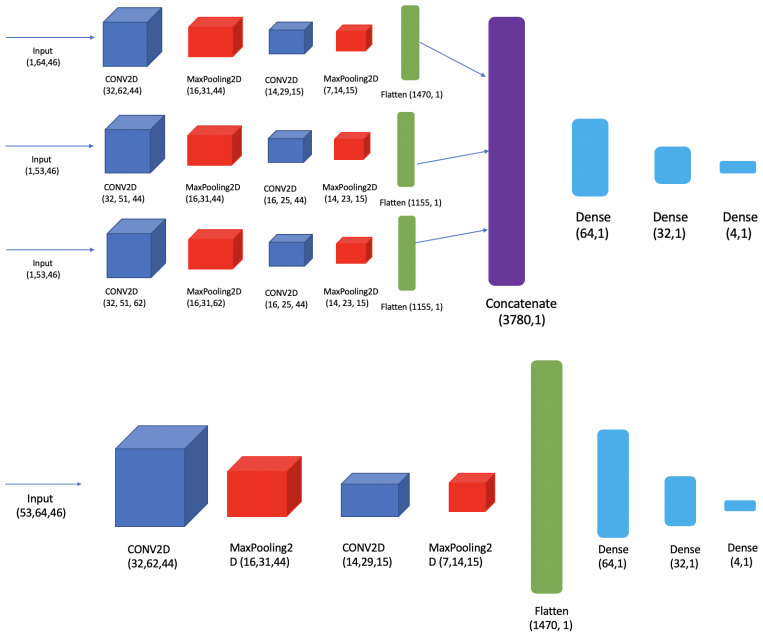

FIG. 5. Architecture of the two models we built. The upper model is the three-feature-extractor model we construct, and the lower model is the classical combination of feature extractor and label classifier.

vector that is passed into a three-layer densely connected network for predicting the labels. The advantage of such an architecture is that splitting the point clouds into three feature extractors are more interpretable since each feature extractor represents fixing an axis of the variance matrices and look at the other two axes. The disadvantage is that the averaging process along each axis may undermine important features globally.

The second model employs a classical combination of feature extractor and label classifier. The first dimension, denoted as x, is fixed as channels, so one slice of the image generated by the other two dimensions at the fixed x is passed into each channel. The model preserves every possible image of the 3D point cloud, which is more computationally expensive. It is expected that the training accuracy could be higher as it preserves everything, but the model may overlook noises as well.

B. Classification With Variance

We train our two models with 3d point clouds that are the *Variance* of each sample. All accuracy results of the classification after 15 epochs are shown in Table. IV B.

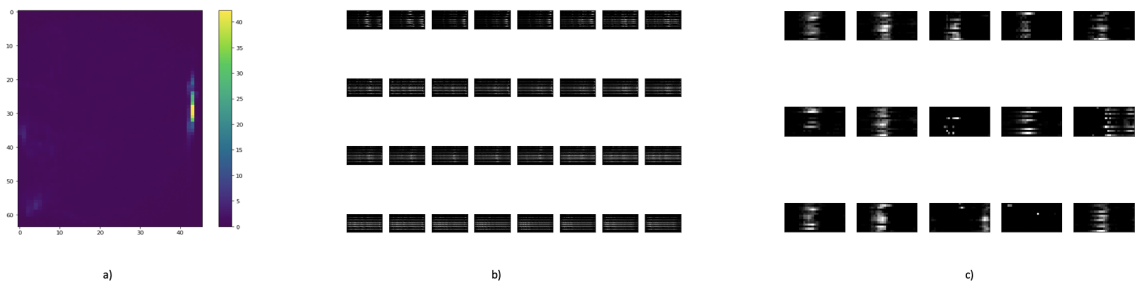

FIG. 6. Feature maps of a selected variance sample. a) The variance sample. As we can see the sample itself is not even in an interpretable domain. b) The feature maps of the sample in the first hidden layer. c) The feature maps of the sample in the second hidden layer.

Model	Training Accuracy	Test Accuracy
Variance Model 1	64.3	62.5
Variance Model 2	63.1	36.8
Mean Model 1	2.38	/
Mean Model 2	27.4	/

TABLE I. Training and Test Accuracy results of the models.

The models perform reasonably well on the training set, with a training accuracy of 64.3 and 62.5 percent. To localize the pattern, we look at feature maps for both high and low-level features, which are shown in Fig. 6. These feature maps are in general not interpretable, as variance itself is not too. Thus, this urges us to train the model with a more interpretable metric, which is the *Arithmetic Mean*, as we will discuss next.

C. Classification with Mean

We train our two models with 3d point clouds that are the *Arithmetic Mean* of each sample. All accuracy results of the classification after 15 epochs are shown in Table. IV B.

As evident in the table, since the mean for each diagnosis type is not very different from each other and we only have few data points, the gradient converges very slowly to optimum: training it with 30 epochs can only bring it up to less than 30%. For the reason, we immediately halt further experiments, as we don't think the model is capable of extracting features from the *Arithmetic Mean* data.

D. Discussion

Comparing the two models trained with Variance, we claim that the model with three feature extractor performs better than the one with one feature extractor. The reason is likely that the variance matrices are themselves noisy, so an averaging process could undermine the noises.

By visualizing the feature maps, we find that the sparsity of activation in the filters is high, which usually leads to high test accuracy. The low test accuracy, in our case, however, implies that the model we trained is highly biased because of the lack of datapoint.

Comparing deep learning models with pixel-wise KL Divergence, we claim that KL Divergence outperforms deep learning models in both neural pattern recognition and pattern localization. The fundamental reason is that our selected dataset of 125 samples are still relatively too small for deep learning models as they have high complexity. In the task of pattern localization, the uninterpretability of deep learning models also makes it hard for us to understand the internal saliencies of the feature extraction, while these internal saliencies are important for determining the location of the neural activity. Besides, it is harder to preprocess data for deep learning than to preprocess data for pixel-wise KL Divergence. For example, when computing pixel-wise KL Divergence, we downsample the pixel distributions to bins with variance at a maximum of 20. It is hard, however, to select these pixels out of the deep learning data, as we fit the whole variance matrices as data for the deep learning models. We propose that a better preprocessing technique may help these deep learning models to perform better.

V. CONCLUSION AND FUTURE WORKS

To conclude, we examine the performance of deep learning models on processing fMRI data from ADHD-200 data comparing with pixel-wise KL Divergence. We first perform phenotypic analysis and data visualization of the fMRI data for the ADHD-200 data. We then define and compute pixel-wise KL Divergence for fMRI data of kids from each diagnosis type, and find that pixel-wise KL Divergence performs reasonably well in both pattern recognition and localization. Finally, we train two deep learning models with both *Arithmetic Mean* and *Variance* data. We find that both models perform much poorer in pattern

recognition and localization than pixel-wise KL Divergence, regardless of the training data.

Future works related to our experiments can benefit from:

1. Establish a better preprocessing technique for training deep learning models with fMRI data.
2. Compare our observations with those made in a larger fMRI dataset.
3. Test the statistical significance of the neural activity pattern and location we found with pixel-wise KL Divergence.

---

[1] Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., Craddock, R. C. (2017). The neuro bureau ADHD-200 preprocessed repository. Neuroimage, 144, 275-286.

[2] Milham, M. P., Fair, D., Mennes, M., Mostofsky, S. H. (2012). The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers in systems neuroscience, 6, 62.

[3] Dai, D., Wang, J., Hua, J., He, H. (2012). Classification of ADHD children through multimodal magnetic resonance imaging. Frontiers in systems neuroscience, 6, 63.

[4] Kuang, D., Guo, X., An, X., Zhao, Y., He, L. (2014, August). Discrimination of ADHD based on fMRI data with deep belief network. In International Conference on Intelligent Computing (pp. 225-232). Springer, Cham.

[5] Kuang, D., He, L. (2014, November). Classification of ADHD with deep learning. In 2014 International Conference on Cloud Computing and Big Data (pp. 27-32). IEEE.