The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Learning Differences between Visual Scanning Patterns Can Disambiguate Bipolar and Unipolar Patients

Jonathan Chung,[1] Moshe Eizenman,[1,2] Uros Rakita,[3] Roger McIntyre,[3,4,5] Peter Giacobbe[3,5]
[1] Department of Electrical and Computer Engineering
[2] Ophthalmology and Vision Sciences & Institute of Biomaterials and Biomedical Engineering
[3] Department of Psychiatry, [4] Department of Pharmacology and Toxicology
University of Toronto, Toronto, ON
[5] Department of Psychiatry, University Health Networks, Toronto, ON

Abstract

Bipolar Disorder (BD) and Major Depressive Disorder (MDD) are two common and debilitating mood disorders. Misdiagnosing BD as MDD is relatively common and the introduction of markers to improve diagnostic accuracy early in the course of the illness has been identified as one of the top unmet needs in the field. In this paper, we present novel methods to differentiate between BD and MDD patients. The methods use deep learning techniques to quantify differences between visual scanning patterns of BD and MDD patients. In the methods, visual scanning patterns that are described by ordered sequences of fixations on emotional faces are encoded into a lower dimensional space and are fed into a long-short term memory recurrent neural network (RNN). Fixation sequences are encoded by three different methods: 1) using semantic regions of interests (RoIs) that are manually defined by experts, 2) using semi-automatically defined grids of RoIs, or 3) using a convolutional neural network (CNN) to automatically extract visual features from saliency maps.

Using data from 47 patients with MDD and 26 patients with BD we showed that using semantic RoIs, the RNN improved the performance of a baseline classifier from an AUC of 0.603 to an AUC of 0.878. Similarly using grid RoIs, the RNN improved the performance of a baseline classifier from an AUC of 0.450 to an AUC of 0.828. The classifier that automatically extracted visual features from saliency maps (a long recurrent convolutional network that is fully data-driven) had an AUC of 0.879. The results of the study suggest that by using RNNs to learn differences between fixation sequences the diagnosis of individual patients with BD or MDD can be disambiguated with high accuracy. Moreover, by using saliency maps and CNN to encode the fixation sequences the method can be fully automated and achieve high accuracy without relying on user expertise and/or manual labelling. When compared with other markers, the performance of the class of classifiers that was introduced in this paper is better than that of detectors that use differences in neural structures, neural activity or cortical hemodynamics to differentiate between BD and MDD patients. The novel use of RNNs to quantify differences between fixation sequences of patients with mood disorders can be easily generalized to studies of other neuropsychological disorders and to other fields such as psychology and advertising.

1 Introduction

Bipolar Disorder (BD) and Major Depressive Disorder (MDD) are two common and debilitating mood disorders. The prevalence of MDD has been estimated to be 16.2% in the United States, affecting between 32 and 35 million adults in their lifetime (Kessler et al. 2003). BD is estimated to affect between 7.5 to 8 million adults in the United States resulting in a lifetime prevalence of between 1 to 5%, however such estimates may be an underestimate due to unaccounted cases of bipolar spectrum disorders (Bauer and Pfennig 2005). Both disorders produce significant functional and cognitive impairment that compromises the quality of life and increases health-care costs (Grande et al. 2016; Simon 2003).

BD is defined by the lifetime presence of symptoms of depression, mania and hypomania. MDD, on the other hand, is only defined by symptoms of depression. However, BD patients spend the majority of the time in clinical depressive states (Judd et al. 2002). Currently, there are no biomarkers or pathognomonic clinical differences to aid clinicians in distinguishing between patients with BD and MDD while in a depressive state (Goodwin and Jamison 2007; Mitchell et al. 2008), increasing the risk of misdiagnosis. In fact, misdiagnosing BD as MDD is relatively common, as revealed by Stensland et al., (2010), who analyzed claims from a United States health plan and found that more than a quarter of BD patients were incorrectly diagnosed at one point during their illness as having MDD. Importantly, misdiagnosis can lead to sub-optimal treatment as MDD is treated with antidepressants, while this class of medication is not recommended as a first-line approach to those with BD. Furthermore, antidepressant can worsen mood cyclicity in those with BD, which can lead to poor outcomes as well as increased medical costs (Bowden 2001) pointing to the need for improved tools to help disambiguate the underlying polarity of the illness in those who present in depressed states.

Differences in emotional and attentional processing in MDD and BD populations is one approach to improve diagnostic precision (Grotegerd et al. 2014). Neuroimaging studies have uncovered important neural mechanisms that differentiate emotional and attentional processing in BD and MDD patients. These discoveries suggest unique traits that may also be present in behavioural responses to emotionally charged themes. Considering that eye movements are reflec-

tive of internal thought processes and closely follow shifts in attention (Kowler 1995; Posner and Dehaene 1994), an alternative experimental paradigm to monitoring attentional biases in patients with mood disorders is to analyze their eye movements using eye-tracking technologies (Eizenman et al. 2003). Unlike neuroimaging technologies, eye tracking technologies can monitor momentary shifts in attentional allocation patterns and are less clinically demanding on patients. Eye-tracking studies have shown that depressed MDD patients fixate on dysphoric images for significantly longer periods of time than healthy controls (Eizenman et al. 2003; Sears et al. 2010; Kellough et al. 2008; Caseras et al. 2007; Duque and Vázquez 2015). In contrast, BD patients fixate less time on happy images than healthy controls but had similar fixation times on sad images (García-Blanco et al. 2014). The above studies point to differences between the visual scanning patterns of the two groups on emotional stimuli.

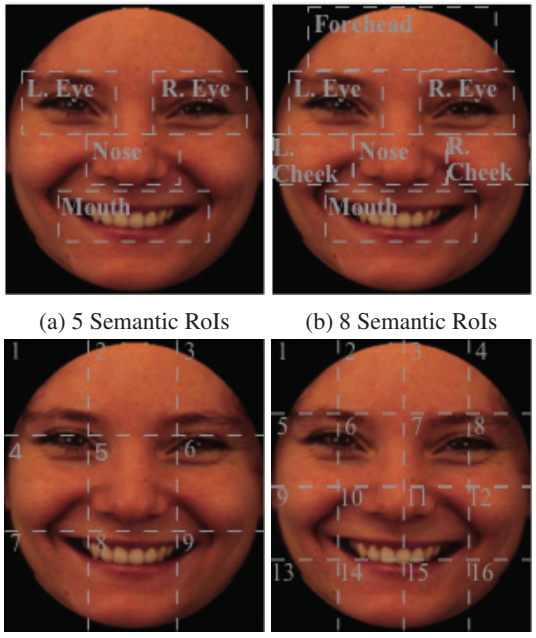
In this paper, we used visual scanning patterns on emotional stimuli to differentiate between BD and MDD patients. More specifically, we used deep learning techniques to capture spatial-temporal differences between sequences of fixations of BD and MDD patients on emotional faces. In the method that we developed, encoded sequences of fixations were fed into a recurrent neural network (RNN) consisting of long-short term memory (LSTM) cells. We used three techniques to encode sequences of fixations on emotional faces. The first two techniques encode sequences of fixations using manually defined semantic or semi-automatically defined grid regions of interests (RoIs) into a series of one-hot encoded vectors. The third technique is a fully data-driven approach that utilises a shallow convolutional neural network (CNN) to extract visual features from the saliency map of individual fixations (long recurrent convolutional network, LRCN (Donahue et al. 2015)).

The first objective of this paper is to develop a classifier that use deep learning techniques to differentiate between fixation sequences of BD and MDD patients on emotional faces. This objective was evaluated by comparing the performance of a baseline logistic regression classifier to that of an RNN classifier. The second objective is to develop a classifier that does not require manual labelling when encoding fixation sequences. To achieve the second objective we explored a fully data-driven deep learning method that use a CNN to extract visual features from saliency maps.

2 RELATED WORK

2.1 Semantic and grid RoIs

Figure 1 displays two methods of determining RoIs for the analysis of visual scanning patterns (Mathôt et al. 2012). Semantic and grid RoIs have been used to assist in the development of tractable handcrafted features from sequences of fixations on visual stimuli. Handcrafted features such as the percentage of fixations within specific RoIs (e.g., eye, nose, mouth) have been used to study human's perception of faces (e.g., (Blais et al. 2008; Weigelt, Koldewyn, and Kanwisher 2012)). Outside the context of face perception, the string edit distance between sequences of fixations within RoIs was used to quantify differences between scanning patterns (e.g., (Cristino et al. 2010; West et al. 2006)). To the best of our knowledge, there are no applications of deep learning to the study of the neuropsychological state of individuals using sequences of fixations within RoIs.



(a) 5 Semantic RoIs (b) 8 Semantic RoIs

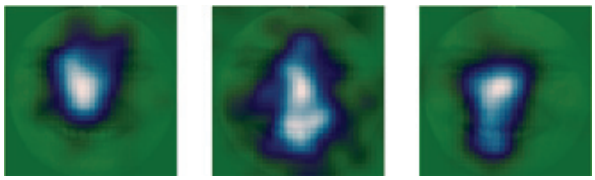
(c) 3×3 Grid (d) 4×4 Grid

Figure 1: Methods of determining RoIs for the analysis of visual scanning patterns: a) 5 Semantic RoIs, b) 8 Semantic RoIs, c) 3×3 Grid, and d) 4×4 Grid

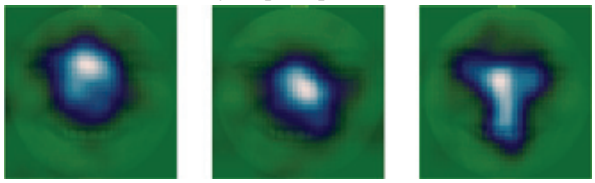
2.2 Saliency maps

Saliency maps have been used to study the spatial distribution of eye fixations on visual stimuli. These studies include anecdotal observations of user interfaces on web pages (Cutrell and Guan 2007) or correlation between saliency maps of different observers (Caldara and Miellet 2011). Although saliency maps have not been studied as inputs to deep learning networks, deep learning models of visual attention have been recently gaining traction. Pan et al. (2016) used two convolutional neural networks (CNNs), a shallow network that was trained end-to-end and a pre-trained deep CNN, to perform saliency prediction. Kruthiventi et al. (2016) described a framework to predict saliency maps and extract regions of salient objects within visual stimuli. In their framework, the saliency maps are predicted using a CNN and a conditional random field on top of the saliency map is used to detect objects within images.

Examples of saliency maps from three patients with BD and three patients with MDD are shown in Figure 2. As CNNs were effective in predicting allocation of visual attention (fixations) in visual stimuli, we trained CNNs to extract visual features that can differentiate between BD and MDD patients from saliency maps.



(a) Saliency maps of patients with BD



(b) Saliency maps of patients with MDD

Figure 2: a) Examples of saliency maps for three BD patients, b) examples of saliency maps for three MDD patients. Note that due to a large inter subject variability between patients in the same group it is difficult to determine a set of spatial features that can differentiate between the saliency maps of the two groups.

2.3 LSTM RNN

We utilise the LSTM architecture to learn differences between fixation sequences of the two groups of patients. The LSTM was introduced to learn the long term dependencies of sequential data (Hochreiter and Schmidhuber 1997). The success of the LSTM (or the similar RNN unit: gated recurrent unit (GRU) (Chung et al. 2014)) has been demonstrated in numerous applications of natural language processing (Kiros et al. 2015; Bowman et al. 2015). Specifically, the Long-term Recurrent Convolutional Network (LRCN) has been demonstrated to learn spatial and temporal interactions of sequential visual inputs. This was described by Donahue et al. (2015) where the authors demonstrated its effectiveness in image and video description generation and activity recognition from sequences of images. Similarly, (Xu et al. 2015) and (Karpathy and Fei-Fei 2015) have demonstrated the effectiveness of the technique in image captioning and image annotations, respectively. In this paper, spatial-temporal interactions within visual scanning patterns on emotional images were explored using the LRCN.

Applications in health care utilising RNNs have also been gaining traction recently. Lipton et al. (2015) used the LSTM to learn a time series of 13 variables (e.g., blood pressures, heart rate, etc.) and to develop a multi-label classification of 128 conditions (e.g., heart failure, seizures, etc.). RNN with GRU cells has been used successfully to predict patient's diagnosis, prescribed medication and the time of their next visit given their visit records (diagnosis, medication, procedure codes, etc.) (Choi et al. 2016). Other recent work compared the capacity of simple RNN, LSTM, and GRU cells to predict the medications of a patient based on billing codes that indicated their conditions and the reasons for the visits (Bajor and Lasko 2016). Recent applications largely focus on incorporating a wide range of input data from different modalities to predict a wide range of diseases, medications, etc. In this work, we demonstrate the advantages of using RNNs to differentiate between patients with two specific disorders that cannot be disambiguated accurately by clinicians.

2.4 Novel contributions

- A novel and accurate detector that uses a deep neural network to quantify differences between ordered fixation sequences of BD and MDD patients in a depressive state. The detector uses manually or semi-automatically defined RoIs to encode visual scanning patterns on emotional faces.
- A fully data-driven detector that uses LRCN to differentiate between BD and MDD patients. The detector does not require user input.

3 Dataset

Seventy-three patients with BD or MDD were tested. All participants were evaluated by a psychiatrist and were between 18 and 65 years old. Patients met the Diagnostic and Statistical Manual of Mental Disorders 5th Edition (DSM-V) criteria (Association and others 2013) for either MDD or BD. All the patients were in a depressed state with scores equal or greater than 20 on the 17-item Hamilton Rating Scale for Depression (HAMD-17) (Hamilton 1960). Twenty-six patients had BD and forty-seven patients had MDD. All participants consented to the study procedures.

Each patient viewed a series of 50 slides. Each slide contained four images placed in a 2×2 configuration. Fifteen slides contained images of emotional faces with happy and sad expressions that were selected from the Karolinska Directed Emotional Faces (KDEF) database (Lundqvist, Flykt, and Öhman 1998). Images of faces provide a good medium to assess the spatial-temporal differences between visual scanning patterns since the images can portray a wide range of emotions and the positions of the facial features can be matched spatially to create repeatable measures. The remaining 35 filler slides were used at the beginning of the test (to allow the participants to get used to the pace of the presentation) and to mask the purpose of the study. The filler slides were not analysed in this study.

Visual attention scanning technology (VAST, EL-MAR Inc. Toronto, Ontario, Canada) (Chau et al. 2015; Pinhas et al. 2014) was used to measure the subjects' eye gaze positions on visual stimuli that were displayed on the VAST's monitor. The eye tracking system in VAST is mounted on a 23 inch LCD monitor and consists of three infrared (IR) light sources, an IR video camera and a processing unit. VAST estimates binocular gaze positions 30 times/sec with an accuracy of 0.5° (Guestrin and Eizenman 2007). During the test, subjects sat approximately 65 centimetres away from the monitor. Following a short eye-tracking calibration procedure, subjects viewed the slides presented on the LCD monitor and their gaze positions were recorded. Each slide was presented for 10.5 seconds and the total test time/subject was less than 10 minutes. The raw gaze positions were segmented into sequences of fixations that were linked to the content of the images on each slide.

4 Methods

We explored three methods to encode the sequences of fixations. In the first two methods, fixations within manually defined semantic or semi-automatically defined grid RoIs were encoded into one-hot vectors. In the third method, the saliency map of each fixation was projected into a lower dimensional feature space using a visual feature extractor (denoted as LRCN). The encoding process reduces the dimensionality of the input features (Choi et al. 2016). The features were then fed into an LSTM network. We first describe the encoding processes that utilise semantic or grid RoIs (Section 4.1), then we describe the visual feature extractor (i.e., the CNN) of the LRCN (Section 4.2) and the LSTM (Section 4.3).

4.1 Encoding fixations sequences using semantic and grid RoIs

Let a fixation be defined by a Cartesian coordinate within an image such that $\{x|x \in \mathbb{R}^2\}$ and a sequence of fixations to be defined as $\boldsymbol{x} = [x^1, x^2 ... x^N]$ where N is the number of fixations per slide. Each fixation x is assigned to one of the M RoIs in the images (the one that the fixation falls within its boundaries). For each of the encoding methods we analyzed two scenarios so that the effects of the user's strategy in determining the RoIs on the performance of the classifier can be analyzed. For the grid method we divided the images to $M = 9$ or $M = 16$ equally sized RoIs (a 3×3 and a 4×4 grids). Using previous studies of fixation patterns on faces (Mathôt et al. 2012), we manually labelled $M = 5$ (right eye, left eye, nose, mouth, background) or $M = 8$ (forehead, right eye, left eye, left cheek, right cheek, nose, mouth, background) semantic RoIs (see Figure 1 for the definitions used in this paper). For the semantic and grid techniques, the encoded visual scanning pattern on each slide $\boldsymbol{y}$ had N one-hot vectors of size M (i.e., $\boldsymbol{y}$ is of size $N \times M$).

4.2 Encoding fixations sequences using saliency maps of fixations and CNN

In previous studies (Borji and Itti 2012; Liu et al. 2015; Zhang and Sclaroff 2013), each fixation was represented by a normal distribution centred around the mean position of the fixation. The saliency map, which is the sum of all the representations of fixations on a slide (shown in Figure 2), is given by $\{\boldsymbol{S}|\boldsymbol{S} \in \mathbb{R}^{40\times40}\}$ (The relatively small saliency map size was chosen to reduce computational costs. Note that we based our saliency map on (Liu et al. 2015) who used map sizes of 50×50):

$$\boldsymbol{S} = \sum_{x^i \in \boldsymbol{x}} \mathcal{N}(x^i, \Sigma) \tag{1}$$

where $\mathcal{N}(\mu, \Sigma)$ is a normal distribution with mean $\mu \in \mathbb{R}^2$ and a covariance matrix $\Sigma = \sigma I_2$, where $\sigma = 5$ pixels which is equivalent to a visual angle of $\pm 0.5°$ (for 65 cm viewing distance).

To encode a sequence of fixations, a sequence of saliency maps of fixations, which is defined by $\boldsymbol{s} = [\mathcal{N}(x^1, \Sigma), \mathcal{N}(x^2, \Sigma) \ldots \mathcal{N}(x^N, \Sigma)]$, is fed through the visual feature extractor ($\phi_V(s)$) (Donahue et al. 2015) to produce a fixed-length vector representation of the fixation sequence, $\boldsymbol{y} = [\boldsymbol{\phi}^1, \boldsymbol{\phi}^2 \ldots \boldsymbol{\phi}^N]$. A shallow 3 layer CNN was used as the visual feature extractor (ϕ_V). The CNN was designed to resemble the shallow network described in (Pan et al. 2016). A pre-trained network from Pan et al. was not used because the iSUN and SALICON datasets that were used for training were incompatible with our data set as they contained natural images and not images of emotional faces. For the same reason, the deep network in (Pan et al. 2016), which was initialised with layers from the VGG net, was also not used. The CNN was adapted to the current dataset, where the number of parameters was significantly reduced (shown in Figure 3). To study the effects of different layers within the CNN on the performance of the LRCN classifier we used ablation experiments (1st layer, 2nd layer and 3rd layer; see Figure 3).

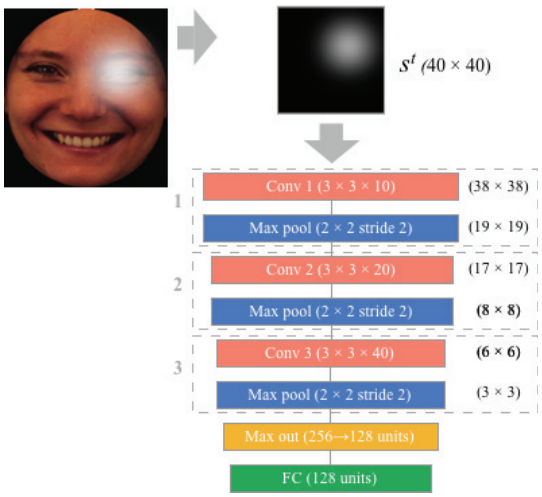


Figure 3: Overview of the visual feature extractor using a shallow 3 layer CNN. Each convolutional and max pooling layers are marked with the image size on the right. The convolutional layers are labelled with the kernel shape × filters, Max pooling layers are labelled with the pool size and stride, and fully connected and max out layers are labelled with the number of units.

4.3 Learning the spatial temporal interactions within fixation sequences with a recurrent neural network

For the three encoding techniques, encoded sequences, $\boldsymbol{y}$, were the inputs to the RNN network. If the length of a sequence of fixations on a slide was smaller than 35, the sequence was zero-padded so the length of the encoded fixation sequences for all the slides and for all the patients were thirty-five (in our study, the maximum number of fixations on a slide was 35).

The RNN utilises a single layer of 128 states LSTM cells

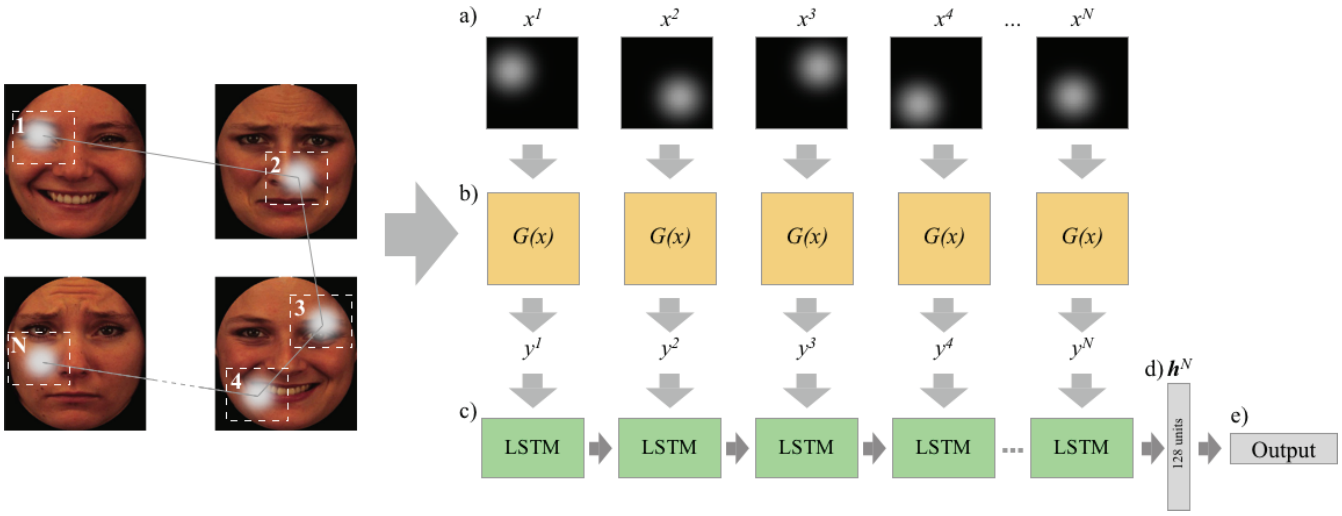


Figure 4: Overview of encoding and learning structures for sequences of fixations. a) A sequence of fixations $\boldsymbol{x} = [x^1, x^2 \dots x^N]$; b) encoders for $\boldsymbol{x}$ (denoted as G); c) LSTM network that learns the spatial-temporal interactions in $\boldsymbol{y}$; d) hidden states $\boldsymbol{h}^N$ that are fed through a fully connected layer to classify the sequence; and e) probability of the patient classification

to extract features from the encoded fixation sequences (Figure 4-c). The network was trained end-to-end. Specifically, let $\boldsymbol{y}^1 \dots \boldsymbol{y}^t \dots \boldsymbol{y}^N$ be each step of the encoded visual sequence (note that for the user defined RoI methods, $G(x)$ is constant). The set of equations for the LSTM (Hochreiter and Schmidhuber 1997) are:

$$\boldsymbol{i}^t = sigm(\boldsymbol{W}_i \boldsymbol{y}^t + \boldsymbol{U}_i \boldsymbol{h}^{t-1} + \boldsymbol{b}_i) \quad (2)$$

$$\boldsymbol{f}^t = sigm(\boldsymbol{W}_f \boldsymbol{y}^t + \boldsymbol{U}_f \boldsymbol{h}^{t-1} + \boldsymbol{b}_f) \quad (3)$$

$$\boldsymbol{o}^t = sigm(\boldsymbol{W}_o \boldsymbol{y}^t + \boldsymbol{U}_o \boldsymbol{h}^{t-1} + \boldsymbol{b}_o) \quad (4)$$

$$\boldsymbol{c}^t = \boldsymbol{f}^t * \boldsymbol{c}^{t-1} + \boldsymbol{i}^t * tanh(\boldsymbol{W}_c \boldsymbol{y}^t + \boldsymbol{U}_c \boldsymbol{h}^{t-1} + \boldsymbol{b}_c) \quad (5)$$

$$\boldsymbol{h}^t = \boldsymbol{o}^t * tanh(\boldsymbol{c}^t) \quad (6)$$

where $\boldsymbol{i}^t$, $\boldsymbol{f}^t$, and $\boldsymbol{o}^t$ are the input, forget, and output gates respectively, $\boldsymbol{c}^t$ is the cell state, $*$ denotes element-wise product, $\boldsymbol{W}$ and $\boldsymbol{U}$ are weights of the LSTM, $\boldsymbol{h}^t$ are the hidden states, and $\boldsymbol{b}_j$ are biases.

At $t = N$, hidden states $\boldsymbol{h}^N$ are fed through a fully connected layer with a softmax activation to predict the classification of the fixation sequence (shown in Figure 4-d).

4.4 Characterising BD and MDD

In our dataset, patients viewed 15 slides and therefore each patient generated 15 independent sequences; thus leading to a multiple-instance learning problem. As the multiple-instance learning problem is not the main focus of this paper, we used a simple approach that takes the mean probability of the individual's fifteen sequences to predict the classification of the fixations sequences for the individual. Specifically, given $\boldsymbol{Y} = \{\boldsymbol{y}^1, \boldsymbol{y}^2 \dots \boldsymbol{y}^k\}$, where $\boldsymbol{y}^i$ represents the encoded sequences for the 15 slides who were viewed by an individual ($k = 15$), the conditional probability for the classification of a patient as BD is given by:

$$P(\boldsymbol{Y}|C = BD) = \frac{1}{k} \sum_{i}^{k} P(\boldsymbol{y}^i|C = BD) \quad (7)$$

where $P(\boldsymbol{y}^i|C = BD)$ is the conditional probability of a BD classification for one of the encoded sequences and is the output of the network (shown in Figure 4-e)

The objective of the network is to minimise the cross entropy for classifying a sequence as BD or MDD.

5 Training and Evaluation method

The networks were implemented with Keras (Chollet 2015) (that was built on top of Tensorflow (Abadi et al. 2016)). For the CNN within the LRCN, the weights of the convolutional layer were initialized uniformly (Glorot and Bengio 2010). After the max out and fully connected layer, dropout layers with a dropout probability of 50% were included. The weights of the LSTM were initialised orthogonally. Dropout layers with a dropout probability of 50% were applied to non-recurrent layers (Lipton et al. 2015). The LSTM was optimised with the Adam algorithm (Kingma and Ba 2014) with a learning rate of 0.001 and mini batch sizes of 40. Early stopping with a patience of 30 was applied.

As a baseline for performance evaluation, we used a simple logistic regression (LR) classifier whose input consists of the number of fixations within each of the user defined ROIs (8 for semantic and 4×4 for the grid RoI encoding techniques). The LR classifier was evaluated with a leave one out cross validation scheme. Because of the limited number of subjects, we employed a leave one out 3-fold cross validation scheme to evaluate the deep learning methods (LSTM with user defined RoIs and LRCN). That is, at each step, all the sequences from one individual were removed and the remaining sequences were fed into a 3-fold cross validation to

randomly split the data into training and validation data. The additional 3-fold cross validation was performed to reduce the chances that the parameters of the model were carefully tuned for the small dataset. The models were trained with 2/3 of the patients in each of the 3-folds until convergence (validation data 1/3 of the subjects) and $P(\boldsymbol{Y}|C = BD)$ (Equation 7) for the left out subject was calculated. The average of $P(\boldsymbol{Y}|C = BD)$ for the 3 folds was calculated for each subject. Our evaluation criteria for the performance of the classifiers when differentiating between patients with BD or MDD were the area under the curve (AUC with 95% confidence intervals (CI) (Kottas, Kuss, and Zapf 2014)) and the balanced accuracy. The balanced accuracy was used to account for the imbalanced dataset and was calculated from the average of the sensitivity and specificity of the classification. The Wilcoxon Rank Sum test was used to determine significant differences between the performance of the classifiers.

6 Results

Table 1 shows that when fixation sequences were encoded by semantic RoIs, the use of RNN improved the performance of the classifier from an AUC of 0.603 [0.466, 0.7408] to an AUC of 0.878 [0.784, 0.972] ($Z = 5.28$, $p < 0.001$). Similarly, when fixation sequences were encoded by grid RoIs, the use of RNN improved the performance of the classifier from an AUC of 0.450 [0.310, 0.589] to an AUC of 0.828 [0.721, 0.937] ($Z = 4.88$, $p < 0.001$). These results suggest that when spatial and temporal interactions within fixation sequences were learnt by a deep neural network, the AUC of the classifier improved for both the grid and semantic methods of encoding by more than 0.25. The performance of the fully automated data-driven method, the LRCN, was 0.879 [0.785, 0.973] which was comparable to that of the classifier that used semantic RoIs (expert system, $Z = 1.14$, $p = 0.254$) and better than the performance of the classifier that used the semi-automatic grid method ($Z = 1.93$, $p = 0.05$). The results show that by using RNN to learn differences between fixation sequences of BD and MDD patients in a depressed state, the diagnosis of individual patients can be disambiguated with high accuracy. Moreover, by using saliency maps and LRCN to encode fixation sequences the method can be fully automated and achieve high accuracy without relying on user defined RoIs.

Table 2 shows the performance of the classifiers as a function of the number of the RoIs and the complexity of the CNN that were used to encode the fixation sequences. The results in Table 2 show that for each encoding technique, the performances of the RNN classifiers are not affected by either the number of RoIs (semantic and grid encoding techniques) or the complexity of the encoder (LRCN encoding technique). The results confirm the robustness of the LSTM architecture.

7 Conclusions

In this paper, we showed that by using RNNs to learn differences between fixation sequences on emotional faces, one can construct a class of classifiers that can accurately differentiate between patients with BD or MDD. The deep learning methods that were presented in this paper improved significantly the performance of a baseline classifier (AUC gain >0.25). The results also suggest that the performance of the RNN classifiers is robust to changes in the parameters of the encoding techniques (number of RoIs, number of layers in the CNN). We further demonstrated that by using CNNs to extract visual features from saliency maps, an automated fully data driven classifier can differentiate between MDD and BD patients with high accuracy (AUC = 0.879).

Method	AUC	Balanced accuracy
LRCN	0.879	0.801
RNN w/ 8 semantic RoIs	0.878	0.821
RNN w/ 4×4 grid of RoIs	0.828	0.773
Baseline Performance		
(a) LR w/ 8 semantic RoIs	0.603	0.581
(b) LR w/ 4×4 grid RoIs	0.450	0.471

Table 1: Classification results.

Method	AUC	Balanced accuracy
LRCN w/ 1 conv. layer	0.869	0.771
LRCN w/ 2 conv. layers	0.872	0.780
LRCN w/ 3 conv. layers	0.879	0.801
RNN w/ 5 semantic RoIs	0.872	0.802
RNN w/ 8 semantic RoIs	0.878	0.821
RNN w/ 3×3 grid RoIs	0.823	0.744
RNN w/ 4×4 grid RoIs	0.828	0.773

Table 2: Model investigations

Several recent studies used neuroimaging techniques to differentiate between MDD and BD patients in the depressed state. Using support vector machines with magnetic resonance images of grey matter volumes, a classifier that was developed by (Redlich et al. 2014) achieved an accuracy of 0.793. Using functional magnetic resonance imaging data Grotegerd et al. developed a binary classifier that achieved a classification accuracy of 0.796 (Grotegerd et al. 2014). Considering that our method outperforms (accuracy=0.821) all of the above methods and that it uses non-invasive remote measurements, can provide results in real-time, is relatively low cost and allows patients to be comfortably tested, it might be more suitable for a clinical setting than any of the methods that were described above.

The method presented in this paper is limited by the small dataset and we could not address the possible confounding effects of types of medication and drug dosages on visual scanning behaviour. With a larger dataset, we could investigate such confounding effects and explore deeper and more sophisticated networks (e.g., more hidden states, larger

saliency map sizes). Also, the information associated with each fixation can be expanded to include higher level knowledge regarding the content of each image or slide (e.g., indicating the expression on the face) to further improve the classification accuracy. Future works could explore methods to learn from fixation sequences on natural scenes, which can lead to a richer set of data to characterise the neuropsychological state of an individual. Finally, the multiple-instance problem that was presented in this paper was not addressed. Methods to account for the dependencies of fixation scanning patterns from the same individual could possibly improve the results of the current study.

Our results suggest that the novel methods described in this study have the potential to help clinicians improve diagnostic accuracy. If validated in studies with larger sample sizes and in multiple sites, this method can lead to improved treatment and outcomes in patients with BD. Furthermore, the novel use of RNNs to quantify differences between fixation sequences of patients with mood disorders can be easily generalized to studies of other neuropsychological disorders and to other fields (e.g., advertising).

Acknowledgments. The study was supported by grant 480479 from the National Science and Engineering Research Council (NSERC) of Canada and an award from the Vision Science Research Program (VSRP) at the University of Toronto. The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Association, A. P., et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Bajor, J. M., and Lasko, T. A. 2016. Predicting medications from diagnostic codes with recurrent neural networks.

Bauer, M., and Pfennig, A. 2005. Epidemiology of bipolar disorders. *Epilepsia* 46(s4):8–13.

Blais, C.; Jack, R. E.; Scheepers, C.; Fiset, D.; and Caldara, R. 2008. Culture shapes how we look at faces. *PloS one* 3(8):e3022.

Borji, A., and Itti, L. 2012. Exploiting local and global patch rarities for saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 478–485. IEEE.

Bowden, C. L. 2001. Strategies to reduce misdiagnosis of bipolar depression. *Psychiatric Services* 52(1):51–55.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; and Bengio, S. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Caldara, R., and Miellet, S. 2011. imap: a novel method for statistical fixation mapping of eye movement data. *Behavior research methods* 43(3):864–878.

Caseras, X.; Garner, M.; Bradley, B. P.; and Mogg, K. 2007. Biases in visual orienting to negative and positive scenes in dysphoria: An eye movement study. *Journal of abnormal psychology* 116(3):491.

Chau, S. A.; Herrmann, N.; Eizenman, M.; Chung, J.; and Lanctôt, K. L. 2015. Exploring visual selective attention towards novel stimuli in alzheimer's disease patients. *Dementia and geriatric cognitive disorders extra* 5(3):492–502.

Choi, E.; Bahadori, M. T.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, 301–318.

Chollet, F. 2015. Keras: Theano-based deep learning library. *Code: https://github. com/fchollet. Documentation: http://keras. io*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cristino, F.; Mathôt, S.; Theeuwes, J.; and Gilchrist, I. D. 2010. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods* 42(3):692–700.

Cutrell, E., and Guan, Z. 2007. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 407–416. ACM.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625–2634.

Duque, A., and Vázquez, C. 2015. Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study. *Journal of Behavior Therapy and Experimental Psychiatry* 46:107–114.

Eizenman, M.; Lawrence, H. Y.; Grupp, L.; Eizenman, E.; Ellenbogen, M.; Gemar, M.; and Levitan, R. D. 2003. A naturalistic visual scanning approach to assess selective attention in major depressive disorder. *Psychiatry research* 118(2):117–128.

García-Blanco, A.; Salmerón, L.; Perea, M.; and Livianos, L. 2014. Attentional biases toward emotional images in the different episodes of bipolar disorder: An eye-tracking study. *Psychiatry research* 215(3):628–633.

Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, 249–256.

Goodwin, F. K., and Jamison, K. R. 2007. *Manic-depressive illness: bipolar disorders and recurrent depression*, volume 1. Oxford University Press.

Grande, I.; Berk, M.; Birmaher, B.; and Vieta, E. 2016. Bipolar disorder. *The Lancet* 387(10027):1561–1572.

Grotegerd, D.; Stuhrmann, A.; Kugel, H.; Schmidt, S.; Redlich, R.; Zwanzger, P.; Rauch, A. V.; Heindel, W.; Zwitserlood, P.; Arolt, V.; et al. 2014. Amygdala excitability to

subliminally presented emotional faces distinguishes unipolar and bipolar depression: an fmri and pattern classification study. *Human brain mapping* 35(7):2995–3007.

Guestrin, E. D., and Eizenman, M. 2007. Remote point-of-gaze estimation with free head movements requiring a single-point calibration. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 4556–4560. IEEE.

Hamilton, M. 1960. A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry* 23(1):56–62.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Judd, L. L.; Akiskal, H. S.; Schettler, P. J.; Endicott, J.; Maser, J.; Solomon, D. A.; Leon, A. C.; Rice, J. A.; and Keller, M. B. 2002. The long-term natural history of the weekly symptomatic status of bipolar i disorder. *Archives of general psychiatry* 59(6):530–537.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Kellough, J. L.; Beevers, C. G.; Ellis, A. J.; and Wells, T. T. 2008. Time course of selective attention in clinically depressed young adults: An eye tracking study. *Behaviour research and therapy* 46(11):1238–1243.

Kessler, R. C.; Berglund, P.; Demler, O.; Jin, R.; Koretz, D.; Merikangas, K. R.; Rush, A. J.; Walters, E. E.; and Wang, P. S. 2003. The epidemiology of major depressive disorder: results from the national comorbidity survey replication (ncs-r). *Jama* 289(23):3095–3105.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, 3294–3302.

Kottas, M.; Kuss, O.; and Zapf, A. 2014. A modified wald interval for the area under the roc curve (auc) in diagnostic case-control studies. *BMC medical research methodology* 14(1):26.

Kowler, E. 1995. Eye movements. *Visual cognition* 2:215–265.

Kruthiventi, S. S.; Gudisa, V.; Dholakiya, J. H.; and Venkatesh Babu, R. 2016. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5781–5790.

Lipton, Z. C.; Kale, D. C.; Elkan, C.; and Wetzell, R. 2015. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*.

Liu, N.; Han, J.; Zhang, D.; Wen, S.; and Liu, T. 2015. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 362–370.

Lundqvist, D.; Flykt, A.; and Öhman, A. 1998. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet* 91–630.

Mathôt, S.; Cristino, F.; Gilchrist, I. D.; and Theeuwes, J. 2012. A simple way to estimate similarity between pairs of eye movement sequences. *Journal of Eye Movement Research* 5(1).

Mitchell, P. B.; Goodwin, G. M.; Johnson, G. F.; and Hirschfeld, R. 2008. Diagnostic guidelines for bipolar depression: a probabilistic approach. *Bipolar disorders* 10(1p2):144–152.

Pan, J.; Sayrol, E.; Giro-i Nieto, X.; McGuinness, K.; and O'Connor, N. E. 2016. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 598–606.

Pinhas, L.; Fok, K.-H.; Chen, A.; Lam, E.; Schachter, R.; Eizenman, O.; Grupp, L.; and Eizenman, M. 2014. Attentional biases to body shape images in adolescents with anorexia nervosa: An exploratory eye-tracking study. *Psychiatry research* 220(1):519–526.

Posner, M. I., and Dehaene, S. 1994. Attentional networks. *Trends in neurosciences* 17(2):75–79.

Redlich, R.; Almeida, J. R.; Grotegerd, D.; Opel, N.; Kugel, H.; Heindel, W.; Arolt, V.; Phillips, M. L.; and Dannlowski, U. 2014. Brain morphometric biomarkers distinguishing unipolar and bipolar depression: a voxel-based morphometry–pattern classification approach. *JAMA psychiatry* 71(11):1222–1230.

Sears, C. R.; Thomas, C. L.; LeHuquet, J. M.; and Johnson, J. C. 2010. Attentional biases in dysphoria: An eye-tracking study of the allocation and disengagement of attention. *Cognition and Emotion* 24(8):1349–1368.

Simon, G. E. 2003. Social and economic burden of mood disorders. *Biological psychiatry* 54(3):208–215.

Stensland, M. D.; Schultz, J. F.; and Frytak, J. R. 2010. Depression diagnoses following the identification of bipolar disorder: costly incongruent diagnoses. *BMC psychiatry* 10(1):39.

Weigelt, S.; Koldewyn, K.; and Kanwisher, N. 2012. Face identity recognition in autism spectrum disorders: a review of behavioral studies. *Neuroscience & Biobehavioral Reviews* 36(3):1060–1084.

West, J. M.; Haake, A. R.; Rozanski, E. P.; and Karn, K. S. 2006. eyepatterns: software for identifying patterns and similarities across fixation sequences. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, 149–154. ACM.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2048–2057.

Zhang, J., and Sclaroff, S. 2013. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, 153–160.