Detection of Apathy in Alzheimer Patients by Analysing Visual Scanning Behaviour with RNNs

Jonathan Chung
University of Toronto
Toronto, ON
jono.chung@mail.utoronto.ca

Sarah A. Chau,
Nathan Herrmann,
Krista L. Lanctôt
Sunnybrook Research Institute &
University of Toronto
Toronto, ON

Moshe Eizenman
University of Toronto
Toronto, ON
eizenm@ecf.utoronto.ca

ABSTRACT

Assessment of apathy in patients with Alzheimer's disease (AD) relies heavily on interviews with caregivers and patients, which can be ambiguous and time consuming. More precise and objective methods of evaluation can better inform treatment decisions. In this study, visual scanning behaviours (VSBs) on emotional and non-emotional stimuli were used to detect apathy in patients with AD. Forty-eight AD patients participated in the study. Sixteen of the patients were apathetic. Patients looked at 48 slides with non-emotional images and 32 slides with emotional images.

We described two methods that use recurrent neural networks (RNNs) to learn differences between the VSBs of apathetic and non-apathetic AD patients. Method 1 uses two separate RNNs to learn group differences between visual scanning sequences on emotional and non-emotional stimuli. The outputs of the RNNs are then combined and used by a logistic regression classifier to characterise patients as either apathetic or non-apathetic. Method 1 achieved an AUC gain of 0.074 compared to a previously presented handcrafted feature method of detecting emotional blunting (AUC handcrafted = 0.646). Method 2 assumes that each individual's "style of scanning" (stereotypical eye movements) is independent of the content of the visual stimuli and uses the "style of scanning" to normalise the individual's VSBs on emotional and non-emotional stimuli. Method 2 uses RNNs in a sequence-to-sequence configuration to learn the individual's "style of scanning". The trained model is then used to create vector representations that contain information on the individual's "style of scanning" (content independent) and her/his VSBs (content dependent) on emotional and non-emotional stimuli. The distance between these vector representations is used by a logistic regression classifier to characterise patients as either apathetic or non-apathetic. Using Method 2 the AUC of the classifier improved to 0.814. The results presented suggest that using RNNs to analyse differences between VSBs on emotional and non-emotional stimuli (a measure of emotional blunting) can improve objective detection of apathy in individual patients with AD.

KDD '18, August 19–23, 2018, London, United Kingdom
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219908

CCS CONCEPTS

• **Computing methodologies → Neural networks**; • **Applied computing → Life and medical sciences**;

KEYWORDS

Recurrent neural networks, Fixation sequences, Representation of visual scanning behaviour, Learning visual scanning styles, Alzheimer's Disease, Apathy

ACM Reference Format:
Jonathan Chung, Sarah A. Chau, Nathan Herrmann, Krista L. Lanctôt, and Moshe Eizenman. 2018. Detection of Apathy in Alzheimer Patients by Analysing Visual Scanning Behaviour with RNNs. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3219819.3219908

1 INTRODUCTION

Apathy is characterised by reduced motivation, social disinterest and emotional blunting, in the absence of mood-related changes, and is the most prevalent neuropsychiatric symptom in patients with Alzheimer's disease (AD). Apathy has been associated with negative effects such as more rapid cognitive and functional decline [21] and a higher risk of mortality [31]. Current methods of assessment rely heavily on the use of clinician's evaluation in corroboration with separate interviews with caregivers and patients [25], which may be ambiguous and time-consuming. More precise and objective methods of evaluation can better inform treatment decisions and prevent the prescription of ineffective or even detrimental therapies. As apathy is one of the risk factors of conversion from mild cognitive impairment to AD, experts in the field have suggested that biomarkers for apathy in the early and pre-symptomatic stage of AD may provide a potential avenue for prevention and early treatment in dementia [11].

Recent evidence supports a link between attention and apathy symptoms [12]. Imaging studies showed that brain regions associated with attention had reduced activity in apathetic compared to non-apathetic AD patients [19]. One component of attention is attentional bias or heightened sensitivity to particular visual stimulus [18]. Eizenman et al., [8] developed a non-verbal objective method to determine attentional biases through the analysis of visual scanning behaviour [28]. In this method, visual scanning behaviour is described by sequences of fixations and saccades within and between regions of interest (ROIs) on visual stimuli [29, 30].

These sequences are affected by both low level (e.g., image saliency) and high level (e.g., memory, emotions) cortical processes [15, 26].

As apathy is associated with social disinterest and emotional blunting, it is reasonable to hypothesise that patients with apathy exhibit smaller changes in their visual scanning behaviours when viewing emotional (e.g., birthday party) and non-emotional (e.g., chair) visual stimuli compared to non-apathetic patients. This hypothesis was tested by Chau et al. [2] who studied differences between fixation times on emotional and non-emotional visual stimuli in apathetic and non-apathetic AD patients. Chau et al. [2] showed that these differences were significantly smaller in apathetic patients. Even though the study by Chau et al. [2] showed that differences in visual scanning behaviour could be used to characterise groups of apathetic and non-apathetic patients, the measure used did not provide an indicator that could reliably differentiate between individual patients.

In this paper, we present two methods that use recurrent neural networks (RNNs) to characterise visual scanning behaviours of apathetic and non-apathetic AD patients. In Method 1, two RNNs were used to learn group-based differences between the visual scanning behaviour on emotional and non-emotional stimuli. The results of the two RNNs were combined to characterise apathy in AD. Method 2 utilises RNNs in a sequence-to-sequence (seq2seq) configuration to model an individual's "style of visual scanning" on non-emotional stimuli. Vector representations of non-emotional and emotional stimuli were then extracted from the model. These vector representations include both content dependent (emotional and non-emotional stimuli) and content independent components of visual scanning behaviour ("style of scanning"). Since the "style of scanning" is consistent for an individual, the effects of the "style of scanning" can be minimised (normalised) by calculating the distance between the vector representations on emotional and non-emotional stimuli (the calculated distance should still maintain differences between components of visual scanning behaviour that are content dependent). The distance between the vector representations is used as a measure of emotional blunting to detect apathy in patients with AD.

2 RELATED WORK

2.1 Modelling visual scanning behaviour with RNNs

Previously, Chung et al. [4] described an RNN structure that was used to learn spatial and temporal information in fixation sequences of patients with bipolar and unipolar disorders. The authors described three approaches to encode sequences of fixations: encoding by grid regions, encoding by semantic ROIs and encoding by convolutional neural network. The encoded sequences of fixation are then fed into RNNs with Long Short-Term Memory (LSTM) cells. Sequences encoded with semantic ROIs achieved the highest classification accuracy between the patient groups. In this paper, we also utilised semantic ROIs definitions where objects within the visual stimuli are manually labelled by domain experts [23]. Specifically, four distinct semantic ROIs are identified for each visual stimuli and sequences were defined based on these ROIs.

2.2 Visual scanning style

Individuals adopt unique styles of visual scanning to obtain information from visual stimuli. These styles, called stereotypical eye movements, are sequences of eye movements that an individual consistently performs while viewing visual stimuli [27]. Some features of the individual's stereotypical eye movements are maintained even when the individual views very different visual stimuli [35, 38]. For example, when an individual views emotional and non-emotional stimuli, the direction of visual scanning (e.g., left to right or right to left) tends to remain the same. Since visual scanning style is less dependent on the content of the stimuli, they behave as nuisance parameters when one analyses the effects of the content of the stimuli on visual scanning behaviour parameters. In this study it is desirable to minimise the effects of nuisance parameters on the measurement of the individual's visual scanning behaviour.

2.3 Extracting vector representations and "style" of data

To extract representations for individuals' visual scanning behaviour, we used neural network vector representation methods that were successfully used in natural language processing applications to create a semantic representation of words [24], sentences [1, 17], and short texts [7]. One advantage of extracting vector representations from neural networks is the ability of the network to learn the "style" of the data. This phenomenon was demonstrated in the context of convolutional neural networks (CNNs) that were able to learn the "styles" of artists [9]. For example, when the art of Van Gogh was used to train a network, the network was able to generate new paintings that were visually similar to the artist's style. In the context of sequential data, Kiros et al. [17] configured RNNs in a seq2seq model and fed a corpus from a specific genre of novels (e.g., romance novels) to the network. The seq2seq models were able to capture the style of the specific genre and generate phrases in the specific style of the genre. We will use RNNs in a seq2seq configuration to learn the individual's visual scanning style.

2.4 Sequence-to-sequence learning

Seq2seq models are used in several natural language processing applications. These applications include machine translation [3, 34], conversations [37], and video captioning [36]. The LSTM or the Gated Recurrent Unit (GRU) architectures were shown to be effective [5] in capturing the relevant information to encode the sequences [34] and learn the specific "style" of the data. To the best of our knowledge, there are no applications of seq2seq learning to eye movement data.

2.5 Novel contributions

- Using RNNs to detect differences between visual scanning behaviours on emotional and non-emotional stimuli to classify apathetic and non-apathetic AD patients.
- Using RNNs to estimate the individual's visual scanning style (stereotypical eye movements).

3 DATASET

Forty-eight patients with AD were tested. Eligibility criteria included a diagnosis of possible or probable AD. Sixteen of the forty-eight patients had a diagnosis of apathy. All participants consented to the study procedures (REB approval was received from the Sunnybrook Health Center, Toronto, Canada).

Each subject viewed a series of 106 slides. Each slide contained four images placed in a 2×2 configuration. Forty-eight slides contained non-emotional stimuli of neutral images (e.g., desk, chair). Thirty-two slides contained emotional stimuli (e.g., social interactions). Twenty-six slides were filler slides that were used at the beginning of the test (to allow subjects to get used to the pace of slide presentation) and to mask the purpose of the study. Images were selected from the International Affective Picture System (IAPS) database [20], based on the images' arousal (feeling of excitement) and valence (feeling of pleasure) ratings. Emotional images had high arousal ratings and either high (social interactions) or low (dysphoric) valence ratings. Neutral images had low arousal ratings and neither high nor low valence ratings.

Visual attention scanning technology (VAST, developed by EL-MAR Inc. Toronto, Ontario, Canada) was used to measure the subject's eye gaze positions on visual stimuli that were displayed on 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° [10, 33]. 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 with a 1-second blank interval between slides.

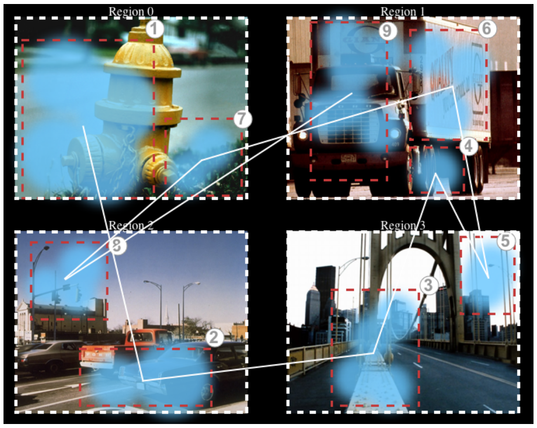

Figure 1: An example of a visual scanning sequence that is defined by a sequence of glances. Blue blobs represent glances (which are outlined with red boxes) and white lines between glances represent saccades. White boxes represent regions of interests within the slide.

The raw gaze position data on each slide is segmented by VAST to a sequence of glances (see Figure 1). Each new glance starts when the subject's gaze moves from one ROI to a new ROI and it ends when the subject's gaze moves away from the new ROI. The minimum duration of a glance is 200 msec. Glance sequences were encoded by the position of the four ROIs on each slide. Each glance $\boldsymbol{g}$ is converted into a one-hot vector of size 5 where the first four elements of the vector correspond to the position of the ROI within the slide (0, 1, 2, and 3 correspond to the four images on the slide, see Figure 1 for the definitions of the ROIs) and the last element of the vector indicates the end of sequence (EOS). The glance sequences were padded with an EOS character so that the length of the glance sequences for all slides were equal (denoted as M_{max}, $M_{max} = 28$ in this study). In this paper, glance sequences are denoted as "visual scanning sequences".

4 METHODS

Section 4.1 describes procedures associated with Method 1 (group differences) and Section 4.2 describes procedures associated with Method 2 (individually normalised differences).

4.1 Method 1: Using group differences between visual scanning sequences on emotional and non-emotional stimuli

Building upon previous work [4], spatial-temporal interactions of visual scanning sequences on emotional and non-emotional stimuli are learnt with an RNN with LSTM cells. Chung et al. [4] fed the hidden states of the last LSTM cell into a fully connected layer to predict the probability that a patient belongs to one of two patient groups. Using this approach two RNNs were trained with visual scanning sequences on non-emotional and emotional stimuli and the mean probabilities that a patient belongs to the apathetic group were calculated for the two types of stimuli (emotional and non-emotional, denoted as $\boldsymbol{\lambda}^g$) (see Figure 2).

Formally, let $\boldsymbol{x} = x_1 \ldots x_t \ldots x_{M_{max}}$ and $\boldsymbol{y} = y_1 \ldots y_t \ldots y_{M_{max}}$ be visual scanning sequences on non-emotional and emotional stimuli respectively. Using sequence $\boldsymbol{x}$ as an example, the elements in $\boldsymbol{x}$ were fed through the following standard set of equations of the LSTM [13].

$$\boldsymbol{i}_t = sigm(\boldsymbol{W}^i \boldsymbol{x}_t + \boldsymbol{U}^i \boldsymbol{h}_{t-1} + \boldsymbol{b}^i) \tag{1}$$

$$\boldsymbol{f}_t = sigm(\boldsymbol{W}^f \boldsymbol{x}_t + \boldsymbol{U}^f \boldsymbol{h}_{t-1} + \boldsymbol{b}^f) \tag{2}$$

$$\boldsymbol{o}_t = sigm(\boldsymbol{W}^o \boldsymbol{x}_t + \boldsymbol{U}^o \boldsymbol{h}_{t-1} + \boldsymbol{b}^o) \tag{3}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * tanh(\boldsymbol{W}^c \boldsymbol{x}_t + \boldsymbol{U}^c \boldsymbol{h}_{t-1} + \boldsymbol{b}^c) \tag{4}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * tanh(\boldsymbol{c}_t) \tag{5}$$

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 each step t, hidden state $\boldsymbol{h}_t$ is calculated from input sequence $x_1 \ldots x_t$ and $\boldsymbol{h}_t$ is then fed into the LSTM cell at $t + 1$. The hidden state $\boldsymbol{h}_{M_{max}}$ is the vector representation of the complete input visual scanning sequence. For training, $\boldsymbol{h}_{M_{max}}$ is fed through a fully connected layer to predict the conditional probability of the sequence ($P(\boldsymbol{x}|C = Apa)$ where C is the classification of the individual and Apa is the apathetic group) originating from an apathetic or a non-apathetic patient with AD.

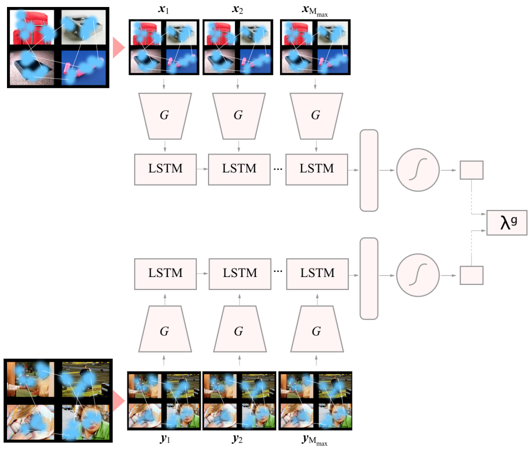

Figure 2: The RNNs of Method 1 learn the spatial-temporal interactions of visual scanning sequences on non-emotional (top) and emotional stimuli (bottom). λ^g is a vector of size 2 containing $P(\boldsymbol{x}|C = Apa)$ and $P(\boldsymbol{y}|C = Apa)$, where $\boldsymbol{x}$ are visual scanning sequences for non-emotional stimuli and $\boldsymbol{y}$ are visual scanning sequences for emotional stimuli, C is the patient classification and Apa is the apathetic group. Function G pre-processes a glance into a one-hot encoded vector.

The networks were trained with a leave-one-out 3-fold cross validation scheme to minimise the cross entropy loss. Specifically, all the sequences from one individual were removed and the remaining sequences from all other subjects were fed into a 3-fold cross validation to randomly split the data into training and validation data. As data from multiple subjects were used during training, we expect this method to learn group differences. The outputs of the networks for the 3-folds were then obtained for the left-out subject and the means were taken for the 3-folds to obtain $P(\boldsymbol{x}|C = Apa)$ for non-emotional and $P(\boldsymbol{y}|C = Apa)$ for emotional stimuli.

Since each individual viewed N non-emotional stimuli, the complete set of visual scanning sequences for non-emotional stimuli is $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_N]$ where each element in the set is a visual scanning sequence on one non-emotional visual stimulus. In a similar manner, each individual viewed N' emotional stimuli and the complete set of visual scanning sequences for emotional stimuli is $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2 \ldots \boldsymbol{y}_{N'}]$, where each element in the set is a visual scanning sequence on one emotional visual stimulus. The individual's classification probabilities, $P(X|C = Apa)$ and $P(Y|C = Apa)$, were calculated by taking the mean of $P(\boldsymbol{x}_k|C = Apa)$ and $P(\boldsymbol{y}_{k'}|C = Apa)$ for $k = [1 \ldots N]$ and $k' = [1 \ldots N']$. Finally, the probabilities $P(X|C = Apa)$ and $P(Y|C = Apa)$ are stored in a vector of size 2 (λ^g - Figure 2) and are used by a logistic classifier.

4.2 Method 2: Using individually normalised differences between the visual scanning behaviour on emotional and non-emotional stimuli

The method described in Section 4.1 uses a single model to describe the differences in spatial-temporal interactions between the two groups of patients and does not take into account the individual's "style of scanning" or his/her stereotypical eye movements. Due to the limited number of slides that each individual viewed, it is difficult to train a model that will take into account the individual's "style of scanning" for each individual. To alleviate such issues, we propose that the RNN described in [4] can be placed in a seq2seq configuration to learn and compensate for an individual's stereotypical eye movements. The method is described in three stages and is illustrated in Figure 3. In the first stage, a seq2seq model learns the individual's "visual scanning style". In the second stage, the trained model creates vector representations that contain information on both the individual's visual scanning style (content independent) and the individual's visual scanning behaviour on non-emotional and emotional visual stimuli (content dependent). In the third stage, the distance between the vector representations of visual scanning sequences on non-emotional and emotional visual stimuli of the individual is calculated. The distance minimises (normalises) information that is common to the vector representations of emotional and non-emotional stumuli. The distance is used as a measure of emotional blunting and is used by a simple logistic regression for classification.

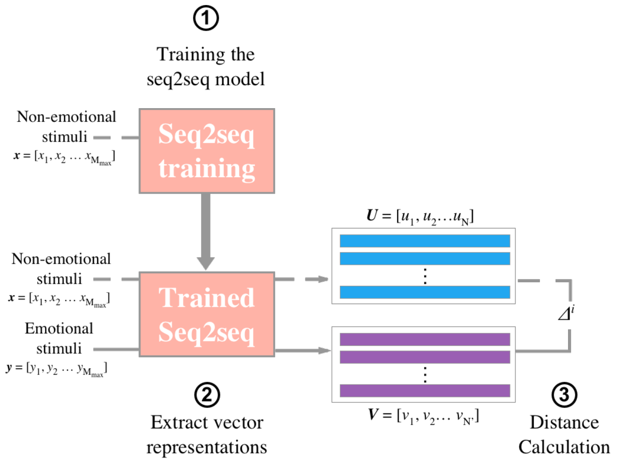
①
Training the seq2seq model
Non-emotional stimuli
$x = [x_1, x_2 \ldots x_{M_{max}}]$
Seq2seq training
Non-emotional stimuli
$x = [x_1, x_2 \ldots x_{M_{max}}]$
Emotional stimuli
$y = [y_1, y_2 \ldots y_{M_{max}}]$
Trained Seq2seq
$U = [u_1, u_2 \ldots u_N]$
Δ^i
$V = [v_1, v_2 \ldots v_{N'}]$
②
Extract vector representations
③
Distance Calculation

Figure 3: The overall architecture of the seq2seq processor of Method 2. N and N' are the number of non-emotional and emotional slides viewed by the individual respectively. Δ^i denotes the distance between the hidden states of an individual.

4.2.1 Training the seq2seq network. In stage 1, our model utilises RNNs in a seq2seq configuration (encoder-decoder) to learn the individual's "style of scanning". The model was trained to encode visual scanning sequences on non-emotional stimuli into hidden states that are used by the model's decoder to reconstruct visual scanning sequences on other non-emotional stimuli (shown in Figure 4). The seq2seq model was trained to learn the "style" of the individual's visual scanning behaviour with images that are of limited interest to the viewer. Since differences between images within non-emotional stimuli have limited effect on the subject, the visual scanning sequences on the non-emotional stimuli are more consistent (easier to learn) and reflect, mainly, the individual's "style of scanning" (stereotypical eye movements).

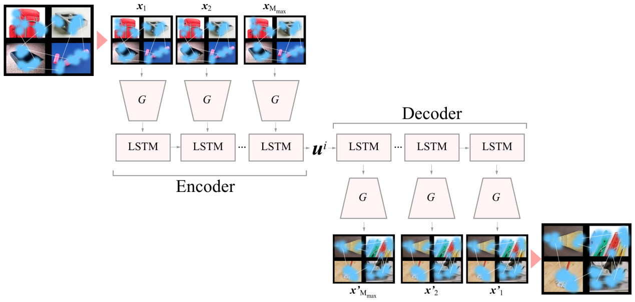

Figure 4: Training the seq2seq model of Method 2: the visual scanning sequence for the input visual stimuli (top) was fed into a series of LSTM cells. The LSTM cells were then used to reconstruct the visual scanning sequence on the output visual stimuli (bottom). Function G pre-processes a glance into a one-hot encoded vector.

Formally, let $\boldsymbol{x} = [x_1 \ldots x_{M_{max}}]$ be visual scanning sequences on a non-emotional stimulus. Following the same procedures described in Section 4.1, the complete input visual scanning sequence $\boldsymbol{x}$ was encoded into hidden states $\boldsymbol{h}_{M_{max}}$ (denoted as $\boldsymbol{u}^i$ where superscript i refers to an individual, see Figure 4). Let $\boldsymbol{x}' = [x'_1 \ldots x'_{M_{max}}]$ be another non-emotional stimulus that was viewed by the same individual where $\boldsymbol{x}'$ is reversed (similar to [34]) and $\boldsymbol{x} \neq \boldsymbol{x}'$. The decoder in Figure 4 is used to generate the output sequence ($\boldsymbol{x}'$). The decoding equations are similar to those of the encoder (Equations 1 - 5), but the initial hidden state is conditioned on $\boldsymbol{u}^i$. In addition, at each glance step t, the hidden state $\boldsymbol{h}_t$ is used to generate the glance x'_t. As such, the network of LSTMs estimates the following conditional probability between the input and output sequences [34]:

$$P(x'_1 \ldots x'_{M_{max}} | x_1 \ldots x_{M_{max}}) = \prod_{t=1}^{M_{max}} P(x'_t | \boldsymbol{u}^i, x'_1 \ldots x'_t) \qquad (6)$$

The objective of the seq2seq model is to minimise the cross entropy between the decoded output sequences and the true output sequences $\{\boldsymbol{x}' | \boldsymbol{x}' \in \boldsymbol{T}, \boldsymbol{x}' \neq \boldsymbol{x}\}$ where $\boldsymbol{T}$ is the training set, given the input $\{\boldsymbol{x} | \boldsymbol{x} \in \boldsymbol{T}\}$. Since this objective will be achieved when the model learns the common aspects of visual scanning sequences on non-emotional stimuli, one can expand the training dataset by permuting through all possible pairs of sequences on non-emotional slides as input and output.

4.2.2 *Extracting and calculating differences between vector representations of visual scanning sequences.*

After training the seq2seq model, the model was used to extract vector representations of visual scanning sequences on non-emotional and emotional stimuli. Sequences $\boldsymbol{x} = [x_1 \ldots x_t \ldots x_{M_{max}}]$ and $\boldsymbol{y} = [y_1 \ldots y_t \ldots y_{M_{max}}]$ were fed into the encoder of the trained seq2seq model, and the hidden states $\boldsymbol{h}_{M_{max}}$ and $\boldsymbol{h}^*_{M_{max}}$ (denoted by $\boldsymbol{u}^i$ and $\boldsymbol{v}^i$ respectively) were extracted.

The normalised difference between the visual scanning behaviour on emotional and non-emotional stimuli of an individual were calculated by subtracting the vector representations $\boldsymbol{v}^i$ and $\boldsymbol{u}^i$ [17]. Kiros et al. [17] demonstrated that differences between the vector representations can be quantified by fitting a linear classifier on top of the subtracted vector representations (i.e., $\boldsymbol{v}^i - \boldsymbol{u}^i$). However, this may be difficult to train with our small dataset. To alleviate this issue, the Euclidean distance between the means of the two vector representations was calculated by $\Delta^i = \| \frac{1}{|\boldsymbol{U}|} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \boldsymbol{u} - \frac{1}{|\boldsymbol{V}|} \sum_{\boldsymbol{v} \in \boldsymbol{V}} \boldsymbol{v} \|$, where $\boldsymbol{U} = [\boldsymbol{u}^i_1 \ldots \boldsymbol{u}^i_N]$ and $V = [\boldsymbol{v}^i_1 \ldots \boldsymbol{v}^i_{N'}]$, where $N = 48$ is the number of non-emotional stimuli and $N' = 32$ is the number of emotional stimuli that were viewed by each individual. Δ^i was fed into a logistic regression for classification.

5 EXPERIMENT DESIGN

5.1 Evaluations

We evaluated the proposed models by determining the classification accuracy of apathetic and non-apathetic patients with AD. We also qualitatively inspected the capacity of the proposed seq2seq model to reconstruct visual scanning sequences of two individuals, and visually evaluated the hidden states extracted from the trained seq2seq model in Method 2.

5.1.1 *Classification accuracy.*

To evaluate the classification accuracy of apathetic and non-apathetic patients with AD, we first presented the results of a baseline classifier that uses the differences between the number of fixations on non-emotional and emotional stimuli [2]. This was compared to the RNN structure of Method 1 (Section 4.1) and the RNN network of Method 2 (Section 4.2).

For the baseline classifier, Method 1 and Method 2 we used the leave-one-out cross-validation scheme to evaluate the classification results. That is, the features of one patient were removed and the logistic regression classifier was trained on the features from the remaining patients to characterise apathetic and non-apathetic patients with AD. The classifier was then applied to the left-out patient and the AUC was calculated with results of the left-out patient. The logistic regression classifier was chosen to demonstrate that the performance of the proposed method to detect apathy could be achieved with a simple classifier.

5.1.2 *Qualitative evaluation of the capacity of the seq2seq model to reconstruct the individual's visual scanning sequences.*

The trained seq2seq models of two patients were used to reconstruct their visual scanning sequences on a held-out set of non-emotional stimuli (i.e., stimuli that were not used in training). We inspected the capacity of the model to capture the "style of visual scanning" of these two individuals.

5.1.3 Visualising the hidden states of non-emotional and emotional stimuli in Method 2. The visual scanning sequences of ten patients were randomly selected from each group for visualisation (a limited number of patients were chosen to reduce the clutter in the plots). The hidden states of the trained RNNs (mean of $\boldsymbol{u}^i$ and $\boldsymbol{v}^i$ in Section 4.1) were extracted from the held out visual scanning sequences and fed into a PCA. The two PCA components with the largest eigenvalues were visualised to observe differences between the hidden states of individuals on emotional and non-emotional stimuli. Also, for the selected individuals the differences between the mean projections of the hidden states on emotional and non-emotional stimuli were visualised. The purpose of visualising these differences is to view how the differences (distance) between the mean projections can be used to reduce the effects of the individual's visual scanning style on the measured visual scanning behaviour and to explore the correlation of the distance with emotional blunting.

5.2 Implementation and training details

The networks consisted of a single layer of LSTM cells. Although the authors in [34] suggested that using multiple layers of LSTM significantly improved the results, a single layer was chosen to reduce the number of weights that have to be learnt as our dataset is limited. The LSTM cells of both Method 1 (described in Section 4.1) and 2 (described in Section 4.2) had 60 hidden states. All the weights were orthogonally initialised. Dropout layers with a dropout probability of 50% were included in the non-recurrent layers of the encoder and decoder LSTM cells [22]. The Adam algorithm with a learning rate of 0.01 [16] was used for optimisation with mini batch sizes of 20. Early stopping with a patience of 30 was applied.

To train the network for Method 2 (Section 4.2), 25% of non-emotional visual stimuli were held out for validating the model for reconstruction. A total of 3840 visual scanning sequences from 48 subjects were analysed (80 slides/subject). The seq2seq network was trained on pairs of visual scanning sequences on non-emotional stimuli i.e., trained on $N^2 - N$ data sequences, where $N = 36$ (75% of the 48 slides of non-emotional stimuli). To train the network from Method 1 (Section 4.1), please refer to [4] for more details.

6 RESULTS

6.1 Classification of apathy

Method	AUC
Handcrafted features [2]	0.646
Method 1 with group differences	0.720
Method 2 with individual differences	0.814

Table 1: Classification results for detecting apathetic and non-apathetic in patients with AD.

Table 1 shows that Method 2 achieved the highest classification accuracy (AUC = 0.814). When the individual's "style of visual scanning" was not used to normalise the visual scanning behaviour (Method 1), the AUC decreased to 0.720. Method 1 achieved an AUC gain of 0.074 compared to the method that used handcrafted features for classification [2]. The results suggest that information embedded in spatial-temporal fixation sequences that describe visual scanning behaviour can provide useful information to improve the classification of apathetic and non-apathetic AD patients.

It is important to note that when the seq2seq model in Method 2 was used with two sets of non-emotional stimuli (in contrast to the method described in this paper that uses one set of emotional stimuli and one set of non-emotional stimuli) and the difference between the hidden states of the two sets was used for an indication for apathy, the AUC was only 0.563. Similar results were obtained when the model was used with two sets of emotional stimuli (AUC = 0.556). Taken together, this results show that only when the differences between the two sets of stimuli reflect emotional blunting (difference between visual scanning behaviour on emotional and non-emotional stimuli), the classifier was able to characterise apathetic from non-apathetic patients with Alzheimer's disease.

6.2 Qualitatively evaluating the capacity of the seq2seq model to learn individuals "styles of scanning"

For the two patients whose models were evaluated, we presented the reconstructed sequence ($\boldsymbol{x}'$ in Figure 4) in the reverse order (denoted as $\boldsymbol{x}'_{rev}$). Specifically, $x'_{rev}(t) = x'(M_{max} + 1 - t)$ so that $\boldsymbol{x}'_{rev}$ is consistent with the progression of time. In addition, most visual scanning sequences of the two patients on non-emotional stimuli had a length of less than 9. That is, the first 19 elements ($M_{max} - 9$, where $M_{max} = 28$ in our study) of $\boldsymbol{x}'_{rev}$ mainly reflect the probability of an EOS characters. Taken together, in Figure 5 we presented the mean probabilities of a glance to one of the ROIs in the reversed reconstructed sequence, $x'_{rev}(t)$ for $t = [20 \ldots M_{max}]$.

We can observe that the two individual models were able to learn several major differences between the sequences of the two patients. The reconstructed sequence of patient 1 suggests that he/she had a fairly consistent sequence length (usually six glances: the probability of EOS is more than 50% in $x'_{rev}(20)$, $x'_{rev}(21)$ and $x'_{rev}(22)$).

During the exploration phases of the visual scanning process (i.e., the first glance to each of the ROIs on the slide), patient 1 had a relatively consistent scanning strategy where the ROI that was viewed first is at the top left of the slide (the mean probability of the first glance to the top left ROI is the highest, see $x'_{rev}(23)$). Since in $x'_{rev}(24)$ the patient moves to the ROIs on the right of the slide, we can infer that his/her main mode of scanning is in a clockwise direction (this observation is supported by the data in $x'_{rev}(25)$ to $x'_{rev}(27)$).

The starting image for patient 2 was less consistent and the sequence length was slightly longer than patient 1 (only the EOS of $x'_{rev}(20)$ is greater than 50%). Initially, patient 2 tended to review the top two images more than the bottom two images ($x'_{rev}(23)$ to $x'_{rev}(25)$). Afterwards, patient 2 had a more random scanning strategy to explore the slide.

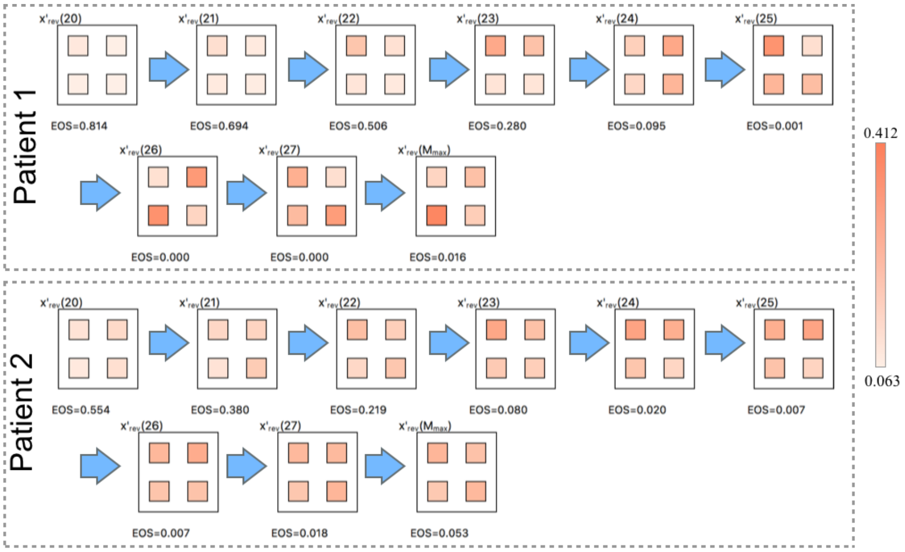

Figure 5: The nine reconstructed glances for two patients. Each reconstructed glance (labelled $x'_{rev}(t)$) contains: the mean probability of a glance on one of the four ROIs on the slides. ROIs are presented by the four squares and the scale that describes the probability of visiting a ROI is shown on the right. The probability of the end of sequence character (EOS) is shown under each step. Note that $x'_{rev}1 - 19$ were omitted as the probablies of EOS were high. Blue arrows shows the progression of time.

6.3 Visualising the vector representation of visual scanning sequences for Method 2

PCA projections of vector representations extracted from the RNNs of Method 2 (Section 4.2) are presented in Figure 6.

Figure 6 shows the first two PCA components of the vector representations of 20 individuals' visual scanning sequences on non-emotional and emotional stimuli. Figure 6 shows clearly that the points from an individual are clustered. The clustering of individuals' projected hidden states suggests that in the vector representations of the visual scanning behaviours in Method 2, the individual's visual scanning style is the largest contributor to the between-subjects data variability. Also, we can observe that the positions of the clusters for each patient are similar for the hidden states on non-emotional and emotional (Figures 6-a and 6-b). This also suggests that an individual's visual scanning style is consistent and independent of the content of the stimuli viewed. For each individual, the projections in Figure 6 may be written as:

$$\boldsymbol{HS}_{non-emo} = VSS + VSB_{non-emo} \tag{7}$$

$$\boldsymbol{HS}_{emo} = VSS + VSB_{emo} \tag{8}$$

where $\boldsymbol{HS}$ is the extracted hidden states vector, VSS is an individual's visual scanning style vector and VSB_{emo} or $VSB_{non-emo}$ are content dependent components of the individuals' visual scanning behaviour.

The results in Figure 6-a and 6-b reveals that there is no obvious separation between apathetic and non-apathetic patients (black versus grey clusters). This is because the large contribution of the individual's visual scanning style to the vector representations of emotional and non-emotional visual scanning behaviours masks differences between the visual scanning behaviours on these stimuli. Since the visual scanning style is independent of the content of the image, the vector representations can be normalised by taking the difference between the hidden states on emotional and non-emotional stimuli.

Figure 7 presents the distance between the individual's mean hidden states on emotional and non-emotional stimuli for the patients whose data were presented in Figure 6 (distance is computed by subtracting the normalised means of the hidden states on emotional and non-emotional stimuli). After the subtraction, the components of the vector that reflect the individuals' "style of scanning" are removed (see Equation 7 and 8) and the effects of emotional blunting on patients in each group have become more apparent. We can observe that the PCA projections for apathetic patients with AD are generally closer to 0, showing that after the subtraction of vector components that reflect the individual's "style of scanning", apathetic patients with AD have similar hidden states for non-emotional and emotional stimuli (an indication of emotional blunting).

7 CONCLUSIONS AND FUTURE DIRECTIONS

Clinical evaluation of apathy in patients with AD is difficult as it relies heavily on subjective interviews with the patient and their caregivers. In this paper, we presented an objective method to detect apathy in AD patients that is based on the analysis of visual scanning behaviour. When the "style of visual scanning" was used to normalise that individual's visual scanning behaviour on emotional and non-emotional stimuli, the method achieved an AUC gain of 0.168 compared to a baseline handcrafted feature classification method [2] The individually normalised method also achieved a gain of 0.094 compared to RNN methods that use group differences

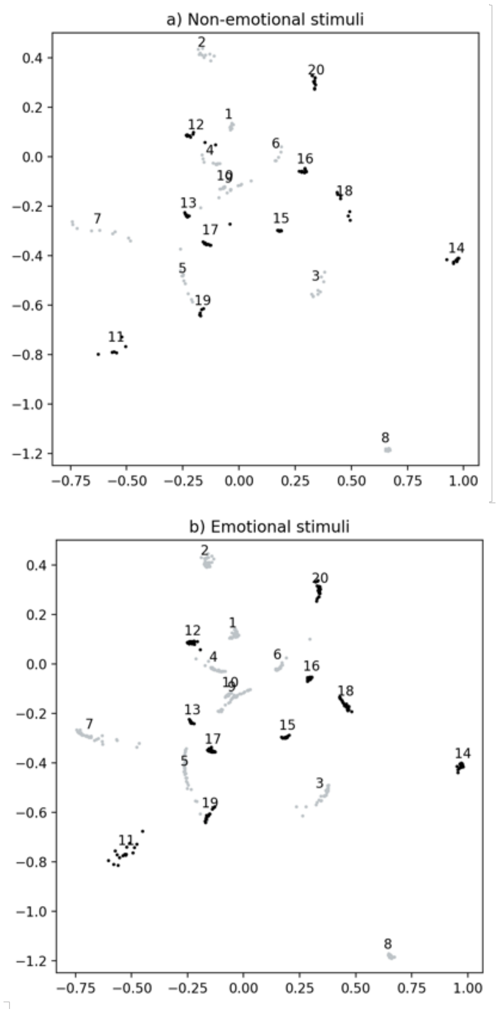

Figure 6: Two-dimensional PCA projections of extracted vector representations (hidden states) of Method 2. The hidden states were extracted from visual scanning sequences on non-emotional (a, 12 slides) and emotional (b, 32 slides) stimuli. Each cluster of points is labelled with a patient number. Black clusters correspond to patients with AD and grey clusters corresponds to apathetic patients with AD.

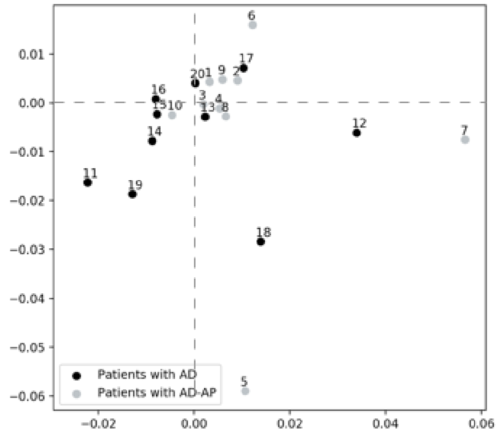

Figure 7: Two-dimensional PCA projections of the differences in the mean hidden states on emotional and non-emotional stimuli of Method 2. Patients with Alzheimer's disease (AD) are indicated in black and apathetic patients with Alzheimer's disease (AD-AP) are indicated in grey. Each cluster of points is labelled with a patient number that corresponds to Figure 6.

between visual scanning behaviour of apathetic and non-apathetic AD patients on emotional and non-emotional stimuli [4].

Several other methods were investigated for the detection of apathetic and non-apathetic patients with AD. Amongst the first works utilised SPECT imaging to characterise apathetic patients with AD [6]. Significant differences were observed in the prefrontal and anterior temporal regions in the brain. More recently, Lanctot et al. [19] used SPECT imaging and observed significant differences in the cerebral blood flow in a number of regions including right orbitofrontal cortex, hippocampus, etc. Similarly, PET imaging revealed a significant decrease in cerebral glucose metabolism in the left orbital frontal regions [14]. MRI imaging revealed that apathetic patients with AD had significantly larger frontal white matter hyperintensity compared to AD patients without apathy [32]. The above neuroimaging techniques primarily focused on differences between group statistics (mean) or correlations with clinical measures. None of the above methods reported individual classification results so it is difficult to compare the results of this study with results of the above neuroimaging studies. As many of the reported studies showed significant differences between apathetic and non-apathetic patients with AD, it will be beneficial to report classification results or effect sizes so that one can determine if these methods can be used to detect apathy in individual AD patients.

In the current study, the Euclidean distance between vector representations of visual scanning behaviours on emotional and non-emotional stimuli was used as an indicator for apathy. In future studies with more subjects, one can use fully connected layers on top of the differences between the vector representations ($\boldsymbol{V} - \boldsymbol{U}$). The fully connected layers can learn appropriate weights and non-linear relationships between the elements of the hidden states to directly learn differences between apathetic and non-apathetic patients with AD.

In general, eye movement studies are carried out with a limited number of subjects who view few visual stimuli (due to time constraints and fatigue). Due to the limited amount of data, researchers typically choose to use handcrafted features rather than machine learning techniques. The method presented in this paper includes several innovations that will enable the use of machine learning techniques in future studies of visual scanning behaviour. These technical innovations include encoding visual scanning behaviour for seq2seq models so vector representations of the visual scanning sequences can be indirectly compared. Indirectly comparing vector representations can normalise individual's visual scanning behaviours so that one can obtain a more sensitive measure of the effects of the content of the visual stimuli on these individual. Finally, expanding the limited dataset in visual scanning studies by permutation can enable and facilitate training of individual models.

ACKNOWLEDGEMENTS

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

[1] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349* (2015).

[2] Sarah A Chau, Jonathan Chung, Nathan Herrmann, Moshe Eizenman, and Krista L Lanctôt. 2016. Apathy and Attentional Biases in Alzheimer's Disease. *Journal of Alzheimer's Disease* Preprint (2016), 1–10.

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[4] Jonathan Chung, Moshe Eizenman, Uros Rakita, Roger McIntyre, and Peter Giacobbe. 2018. Learning Differences between Visual Scanning Patterns can Disambiguate Bipolar and Unipolar Patients. In *Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, Louisiana, USA.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[6] Ann H Craig, Jeffrey L Cummings, Lynn Fairbanks, Laurent Itti, Bruce L Miller, Jenny Li, and Ismael Mena. 1996. Cerebral blood flow correlates of apathy in Alzheimer disease. *Archives of Neurology* 53, 11 (1996), 1116–1120.

[7] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters* 80 (2016), 150–156.

[8] Moshe Eizenman, H Yu Lawrence, Larry Grupp, Erez Eizenman, Mark Ellenbogen, Michael Gemar, and Robert D Levitan. 2003. A naturalistic visual scanning approach to assess selective attention in major depressive disorder. *Psychiatry research* 118, 2 (2003), 117–128.

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[10] Elias Daniel Guestrin and Moshe Eizenman. 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*. IEEE, 4556–4560.

[11] Harald Hampel, Richard Frank, Karl Broich, Stefan J Teipel, Russell G Katz, John Hardy, Karl Herholz, Arun LW Bokde, Frank Jessen, Yvonne C Hoessler, et al. 2010. Biomarkers for Alzheimer's disease: academic, industry and regulatory perspectives. *Nature Reviews Drug Discovery* 9, 7 (2010), 560–574.

[12] Nathan Herrmann, Lana S Rothenburg, Sandra E Black, Michelle Ryan, Barbara A Liu, Usoa E Busto, and Krista L Lanctôt. 2008. Methylphenidate for the treatment of apathy in Alzheimer disease: prediction of response using dextroamphetamine challenge. *Journal of clinical psychopharmacology* 28, 3 (2008), 296–301.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Vjera A Holthoff, Bettina Beuthien-Baumann, Elke Kalbe, Susanne Lüdecke, Olaf Lenz, Gerhard Zündorf, Sebastian Spirling, Kristin Schierz, Peter Winiecki, Sandro Sorbi, et al. 2005. Regional cerebral metabolism in early AlzheimerâĂŹs disease with clinically significant apathy or depression. *Biological psychiatry* 57, 4 (2005), 412–421.

[15] Kai Kaspar, Teresa-Maria Hloucal, Jürgen Kriz, Sonja Canzler, Ricardo Ramos Gameiro, Vanessa Krapp, and Peter König. 2013. Emotions' impact on viewing behavior under natural conditions. *PloS one* 8, 1 (2013), e52737.

[16] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. 3294–3302.

[18] Ernst HW Koster, Rudi De Raedt, Ellen Goeleven, Erik Franck, and Geert Crombez. 2005. Mood-congruent attentional bias in dysphoria: maintained attention to and impaired disengagement from negative information. *Emotion* 5, 4 (2005), 446.

[19] Krista L Lanctôt, Shehnaz Moosa, Nathan Herrmann, Farrell S Leibovitch, Lana Rothenburg, Adolfo Cotter, and Sandra E Black. 2007. A SPECT study of apathy in Alzheimer's disease. *Dementia and geriatric cognitive disorders* 24, 1 (2007), 65–72.

[20] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. 2008. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical report A-8* (2008).

[21] L Lechowski, M Benoit, P Chassagne, I Vedel, D Tortrat, L Teillet, and B Vellas. 2009. Persistent apathy in Alzheimer's disease as an independent factor of rapid functional decline: the REAL longitudinal cohort study. *International journal of geriatric psychiatry* 24, 4 (2009), 341–346.

[22] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. 2015. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* (2015).

[23] Sebastiaan Mathôt, Filipe Cristino, Iain D Gilchrist, and Jan Theeuwes. 2012. A simple way to estimate similarity between pairs of eye movement sequences. *Journal of Eye Movement Research* 5, 1 (2012).

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[25] E Mulin, E Leone, K Dujardin, M Delliaux, A Leentjens, F Nobili, B Dessi, O Tible, L Agüera-Ortiz, RS Osorio, et al. 2011. Diagnostic criteria for apathy in clinical practice. *International journal of geriatric psychiatry* 26, 2 (2011), 158–165.

[26] Vidhya Navalpakkam and Laurent Itti. 2005. Modeling the influence of task on attention. *Vision research* 45, 2 (2005), 205–231.

[27] David Noton and Lawrence Stark. 1971. Scanpaths in eye movements during pattern perception. *Science* 171, 3968 (1971), 308–311.

[28] Leora Pinhas, Kai-Ho Fok, Anna Chen, Eileen Lam, Reva Schachter, Oren Eizenman, Larry Grupp, and Moshe Eizenman. 2014. Attentional biases to body shape images in adolescents with anorexia nervosa: An exploratory eye-tracking study. *Psychiatry research* 220, 1 (2014), 519–526.

[29] Daniel C Richardson and Michael J Spivey. 2004. Eye tracking: Characteristics and methods. *Encyclopedia of biomaterials and biomedical engineering* (2004), 568–572.

[30] Daniel C Richardson and Michael J Spivey. 2004. Eye tracking: Research areas and applications. *Encyclopedia of biomaterials and biomedical engineering* (2004), 573–582.

[31] Gianfranco Spalletta, Jeffrey D Long, Robert G Robinson, Alberto Trequattrini, Sonia Pizzoli, Carlo Caltagirone, and Maria D Orfei. 2015. Longitudinal neuropsychiatric predictors of death in Alzheimer's disease. *Journal of Alzheimer's Disease* 48, 3 (2015), 627–636.

[32] Sergio E Starkstein, Romina Mizrahi, Aristides A Capizzano, Laura Acion, Simone Brockman, and Brian D Power. 2009. Neuroimaging correlates of apathy and depression in AlzheimerâĂŹs disease. *The Journal of neuropsychiatry and clinical neurosciences* 21, 3 (2009), 259–265.

[33] Veit Sturm, Daniel Cassel, and Moshe Eizenman. 2011. Objective estimation of visual acuity with preferential looking. *Investigative ophthalmology & visual science* 52, 2 (2011), 708–713.

[34] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[35] Geoffrey Underwood, Tom Foulsham, and Katherine Humphrey. 2009. Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. *Visual Cognition* 17, 6-7 (2009), 812–834.

[36] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*. 4534–4542.

[37] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[38] Alfred L Yarbus. 1967. *Eye movements during perception of complex objects*. Springer.