Automatic Detection of Mind Wandering Using Residual Network Generated Features

Arbër Demi
Supervisor(s): Bernd Dudzik, Xucong Zhang, Hayley Hung

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 19, 2022

Abstract

Mind wandering occurs when a person’s attention unintentionally shifts away from their current thought or task. Being able to automatically detect cases of mind wandering can assist applications with attention retention, and help people with maintaining focus. Many methods have been tested to deal with mind-wandering detection, but they are mainly conducted in controlled environments. There also has been little study into the usefulness of learned features from neural networks. This paper is focused on showcasing the effectiveness of using neural network generated features as input for classification models. Specifically, using ResNet to generate features which are then used as input by supervised learning models for classification. These features and models were used to classify mind wandering in the Mementos data set, outside of a controlled environment or differently put as “In-the-wild“. The study shows that the extracted features could not be used to accurately detect mind wandering based on the F1-Score (Macro) measure. The results can be attributed to data imbalance, low amount of data, lack of dataset-tailored pre-processing operations, and indiscriminate features. To improve on the study, more data collection is advised and the usage of methods like resampling and data augmentation to deal with data imbalance. And lastly, experimentation with neural network training and transforming the data into a time series format to better represent the temporal information from the data.

1 Introduction

Mind wandering has many definitions based on the context it is being studied under. The definition used for this research is quite similar to the one described by *Smallwood and Schooler, 2006* and aligns with *Schooler et al., 2014*: “When mind wandering occurs, the executive components of attention appear to shift away from the primary task, not due to external factors or the person interacting with the external environment“.

Mind wandering is an important field to study because there is a growing interest in enabling intelligent applications to automatically detect episodes of mind wandering in their users, providing an opportunity for these applications to take action.

This study is focused on the automatic detection of mind wandering “In-the-wild“, meaning in an uncontrolled environment with many varying factors, through the use of neural networks.

Although this specific research question is individual work, it is part of a wider study on automation-detection of mind wandering with other peers in a research group.

1.1 Neural networks and ResNet

A major reason for neural networks being chosen for feature extraction rather than some other feature extraction methods that might focus only on a select few features (like gaze features or facial expressions), is that the end-to-end training of neural networks could provide a larger array of useful distinguishing features that may be missed while focusing on only a select few features.

Another thing to consider is deep learning’s power and success with Computer Vision tasks over previous state-of-the-art machine learning techniques in recent years as mentioned in *Voulodimos et al., 2018*, makes neural networks strong tools to consider for any task in Computer Vision.

For this task, the network used is ResNet. Below is a brief description of what a residual network (ResNet) is:

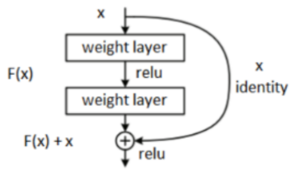


Figure 1. Residual learning: a building block.

A residual network is a network created to solve issues with vanishing gradients in very deep CNNs (Convolutional Neural Networks). As can be seen in Figure 1, this is done through “skip connections“ which can skip over some layers. Including these skip layers allows for information that can be lost through backpropagation in very deep networks to remain, allowing for an increase in performance for deep networks.

This information being no longer lost is what makes ResNet a very robust network to use, and it has been proven to perform very well in Computer Vision tasks ever since it was introduced *He et al., 2016*.

On top of that, ResNet has different versions with different depths to choose from. Using a lower layer version to test out the performance can serve as a good baseline while the higher layer versions can be used for larger databases with better resources for tuning and tweaking values to get maximum performance from the network.

1.2 Related work

Although there have been other studies into automatically detecting mind wandering, some of them have constraints in what they do and how the data was collected.

Works such as *Zhao et al., 2017* and *Steward et al., 2017* although working on mind wandering detection, are focused on specific features and the data was collected in a lab environment. Both of these studies were with supervised learning methods.

Although work similar to *Hosseini and Guo, 2019* shows the use of deep learning to detect mind wandering, it is done with EEG signal data. Nevertheless, the paper can provide insight into what the process is like.

One paper with a topic similar to this research is *Singha, 2021*. The main differences between what is described in the paper and this research are that the data used by Singha is collected in a controlled environment, the features the network is being trained with are only gaze-based ones and the paper focuses on convolutional neural networks, as is explained in the entry of the paper and the data set section in the methodology.

As can be seen by the differences just described, work that is done in the field of mind wandering has some areas still left to be explored.

This study will provide insight on the possible benefits of using a residual network and its generated features to detect mind wandering through the Mementos data set, “**the first multi-modal corpus for computational modeling of affect and memory processing in response to video content**” *Dudzik et al., 2021.*

This data set consists of 1995 individual responses from 297 unique participants reacting to 42 different segments of music videos. The task presented to them was to watch the videos. There will be more information about the usage of the data set in the following section.

Specifically, this research focuses more on the “In-the-wild“ environment and the possibility of more accurate detection through multiple features and a residual network.

1.3 Research question

The main question for this research is: How to use the features generated by a residual network to automatically detect mind wandering through the use of supervised learning models?

The main question can be broken down into several subquestions:

- What is a good definition of mind wandering, considering the context of the dataset and what the neural network should detect?
- What pre-processing operations can be applied to the input data to assist the network in generating more useful features?
- Can the supervised learning models using the generated features achieve better results than a majority class classifier?
- Which of the supervised learning models performs the best when it comes to detecting mind wandering?

With this research, more information will be brought out on what a residual network and its learned features can do to help with the task of automatic mind wandering detection in non-supervised environments or “In-the-wild“.

2 Method and approach

The neural network that is going to be used for this study is ResNet-18. This will be the baseline network to extract features with as a lightweight pre-trained network.

2.1 Data annotation

There were many factors to consider when it comes to data annotation. The data set used is not annotated even though there are self-reports of when mind wandering happens, but for the purpose of this study the self-reports are ignored as the mind-wandering detection should be based on the multi-modal data.

Alongside the other peers, a “rulebook” was created to have objective measures of when mind wandering occurred. At first multiple signs were identified: smile, looking up/rolling eyes, squinting eyes, sounds of person, frown. All of these signs were given extensive descriptions that were agreed upon by all peers. Table 1 shows these signs and their descriptions.

Table 1: 5 different signs are shown all with their corresponding descriptions. This ”rulebook” was used as a general direction for annotating mind wandering cases.

Sign	Description
Smile	Sometimes a smile can be an indication of good memories, so if the smile is very expressive and sudden/genuine smile, it could be a reaction or a response to the video. A very subtle smile could also be a form of reminiscing / remembering a memory so this is also considered a form of mind wandering.
Looking up/eye-rolling	Looking up or rolling eyes are interpreted as looking up for a continuous-time which could be followed by a movement of gaze to the side. Usually, this is caused by an individual trying to remember/recollect.
Eye squinting	Can indicate that person is having a focused thought process happening, which is most likely unrelated to the task of watching the music video.
Sounds from participant	When an individual is speaking to himself, it could indicate that person is going through a thought process and most likely it could be interpreted as mind wandering.
Frowning	Sometimes frowning can be an indication of bad or sad memories, so if the frown is very expressive and sudden/genuine, it could be a reaction or a response to the video. A very subtle frown could also be a form of reminiscing / remembering a memory so this is also considered a possible episode of mind wandering

Following these descriptions and signs, the data was annotated in alternating groups of 3 and 2 people, with unclear cases decided upon with all 5 peers present. This allowed for a faster annotating process while maintaining an unbiased annotating process.

The annotating process consisted of identifying a moment in time where the participant is mind wandering and marking it with a time segment annotation from 1 to 1.5 seconds before the moment that was decided as the beginning of mind wandering, up until 2 to 5 seconds after.

Starting the segment before the moment of mind wandering allows for the annotated segment to have enough data for the models to train properly with a sufficient change in features, while the remainder of the duration varies on the expression of the participant.

To further clarify, some participants showed signs of mind wandering with short lengths, while others had longer-lasting signs. Due to this, a varying length of mind wandering segments is introduced.

Only part of the data set has been annotated due to the time constraint, with 549 total samples from the 1995 available ones. Some data has been deemed inappropriate for the study, as the participants were walking away from the camera or falling asleep. Of the 549 total samples, 435 are valid videos, with 52 having at least one annotation of mind wandering.

2.2 Data imbalance

As can be seen from subsection 2.1, the amount of data annotated as mind wandering is disproportionate to the total amount of samples annotated. This data imbalance is seen quite often in research regarding mind wandering, and it is present in all of the related work that was discussed in this study, which I will mention again for ease of reference: *Steward et al., 2017*, *Singha, 2021*, *Hosseini and Guo, 2019*, *Zhao et al., 2017*.

This problem is usually handled with data augmentation, but experimentation with different augmentation techniques was not done due to the large amount of time that data extraction took. Due to the time constraints of the study, there was not much time for this experimentation.

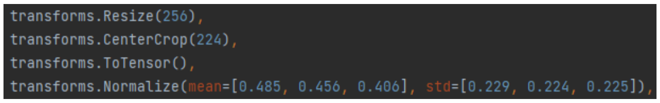
Meanwhile, techniques like SMOTE do not perform very well on high-dimensional data as mentioned *Blagus and Lusa, 2013*, so even though it can be used on extracted features to combat data imbalance, it is not appropriate for this case due to the shape of the data which is talked about in subsection 2.3.

2.3 Data preparation

The videos were split into frames which were used as input for the network. Since the videos are in 30 frames per second, to reduce the amount of unnecessary data, every other frame is skipped, meaning that from 1 second of video we get 15 images. This decision for 15 frames was made after considering the common use in other studies related to human action recognition, as mentioned in *Van Gool, 2008*.

A few frames above what is mentioned in the referenced paper are included since this study is not specifically about recognizing the actions of the participants of the studies, some actions are related more to mind wandering than they are to non mind wandering.

Afterward, the images were then transformed into the format that was expected by ResNet-18 to obtain the generated features that the network would usually use for classification. This transformation is shown in Figure 2:



```
transforms.Resize(256),
transforms.CenterCrop(224),
transforms.ToTensor(),
transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]),
```

Figure 2: The figure shows the code used for the preprocessing of the images, firstly resizing to a 256x256 format, then cropping from the center to a 224x224 format and after turning the image into a tensor which is an n-dimensional array, it is normalized based on the parameters used by ImageNet.

After these changes, the images were put into the network where tensors were extracted. These tensors are of shape 512x7x7.

2.4 Input of the supervised learning models

The tensors were calculated in batches of 100 frames or less in cases where less than 100 frames were remaining from a video. Afterward, the mean of each batch of tensors was calculated and then used as input for the supervised learning model to continue with the classification task.

Experimentation was done with both the 512x7x7 tensors and also 512x1x1 tensors, which were obtained by applying an *AveragePooling2D* operation on the tensors. This is the same as what ResNet itself would use for classification.

Experimentation was done with both shapes of tensors to see which retained more discriminative information to help distinguish between mind wandering and non mind wandering cases, more information about the results of this experimentation can be found in subsection 3.2.

2.5 Different model experimentation

When it comes to different models used for classification, there were several common ones used to check performance results:

1. Random Forest
2. KNN
3. Weighted SVM
4. Decision Tree
5. MLP
6. Stacked
7. Majority

This wide range of classifiers was selected to make sure that any issues in performance were not due to the usage of the wrong classifier. More on the results of experimentation with these models and their results in the following section.

3 Experimental Setup and Results

For the data annotation process, VGG Image Annotator[1] was used.

For the experimentation done in this paper, the latest version of PyTorch was used. Keras cannot be used to reproduce the results of this paper as it does not have an official implementation of ResNet-18 yet, although there are some unofficial implementations. Regardless, something similar can be reproduced with the higher-layer number versions of ResNet.

[1] https://www.robots.ox.ac.uk/vgg/software/via/

For the classifiers, the models provided by sklearn were used. Sklearn was also used for the metrics and parameter tuning.

Several different metrics were used to evaluate the performance of the classifiers used. Specifically, these were accuracy, MCC(Matthews Correlation Coefficient), and F1-Score. F1-Score was calculated using 3 different methods as they are defined in sklearn: Macro, Micro, and Weighted.

Only the Macro version of F1-Score was used in the tables you will find below, as it gives a score that doesn't take label imbalance into account when it is calculating the F1-Score. This means that in a case where one class is being correctly classified all the time while the other is being incorrectly classified all the time, the F1-Score would 0.5. This is useful for this study since the focus is on being able to detect mind wandering cases, not as much on the non mind wandering cases.

The data was split into a 25% test and a 75% training amount. The split here is important as there is not much training data, therefore the model needs a good amount to use for training. The split was based on common uses of the 80% training and 20% split for machine learning.

The data was manually split to make sure that different video responses from the participants were not in different splits. This was done to make sure that no bias was introduced to the model during the training process, as it could happen that 6 out of 7 responses from one participant could be in the training data, with one being in the test data. Then, having encountered data that is similar beforehand, the model could properly label the case in the test data even though it is not a very robust or generally accurate system.

3.1 Comparison of different feature shapes

First, experimentation was done with the feature shapes. As mentioned in sub-section 2.3, the originally extracted features were tensors of shape 512x7x7. These features were extracted instead of the pooled features directly to experiment with the usefulness of the extra information in the 7x7 shaped matrices.

After this extraction, two versions of the data were created. One version had the 512x7x7 shape and one with the 512x1x1 shape. After the mean of the batches of tensors were calculated as mentioned in sub-section 2.4, then this data was separately used to see which was more useful in discriminating between mind wandering and non mind wandering.

From a first look at the labels predicted from both versions of the available data, both types of data seemed to have some issues regardless of the classifier. Mainly, it seemed that they weren't discriminative enough to detect a single case of mind wandering, as every result was coming out as non mind wandering for the test cases. This is also shown by the results from the tables below, one for the pooled data and one for the non-pooled data:

Table 2: Results of the classifiers (pre-tuning for SVM) on the pooled data from the test set

Classifier	Accuracy	F1-Score(macro)	MCC
Random Forest	0.9843	0.4960	-0.007
KNN	0.9153	0.4779	-0.02
Weighted SVM	0.9887	0.4971	0
Decision Tree	0.9674	0.4917	-0.015
MLP	0.9887	0.4971	0
Stacked	0.9887	0.4971	0
Majority	0.9887	0.4971	0

Table 3: Results of the classifiers (pre-tuning for SVM) on the pooled data from the test set

Classifier	Accuracy	F1-Score(macro)	MCC
Random Forest	0.9702	0.4924	-0.0147
KNN	0.8918	0.4714	-0.0348
Weighted SVM	0.9887	0.4971	0
Decision Tree	0.9484	0.4867	-0.0218
MLP	0.8419	0.4570	-0.0441
Stacked	0.8419	0.4570	-0.0441
Majority	0.9887	0.4971	0

These results are not very useful, as all the models seem unable to detect mind wandering cases. Even through manual inspection, there were only 2 instances of mind wandering being detected correctly in a prediction from a model.

There also does not seem to be any major difference between using the pooled and non-pooled data, but using the pooled data did make the models label almost everything as non mind wandering more often than using the non-pooled data.

Since the difference was very small, both types of data were used for the final evaluation.

3.2 Comparison of different models used

The results from the previous sub-section were quite low. This could be due to the features used not being discriminative enough, but the data imbalance problem is likely also a major factor, especially considering the minor amount of mind wandering cases in the test split.

This doesn't provide much help in choosing the proper classification model for this study, but the Weighted SVM was chosen to continue the evaluation, as it is more commonly used for unbalanced data sets, and it also had the best performance in the training split, being the only one to achieve accurate labeling for all data points. This was not included in tables 2 and 3 as the training split is not usually very indicative of a well-performing model.

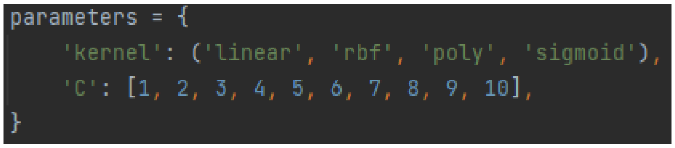
Regardless, considering that the results from the test split were not very informative, this is another small reason to choose the SVM over the other models.

3.3 Tuning of SVM parameters

Due to the high dimensionality in the data and there not being any support from sklearn to run the algorithms used on a

GPU, tuning was a very lengthy process. Due to this, there was not enough time to tune multiple models, therefore only tuning for the Weighted SVM was done.

Tuning was done with the *GridSearchCV* algorithm from sklearn, with the following parameters and values:



```
parameters = {
    'kernel': ('linear', 'rbf', 'poly', 'sigmoid'),
    'C': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
}
```

Figure 3: The figure shows the options for the kernels and the C values used. Most of the kernels available from sklearn were used to check which were most appropriate for the specific data, and only higher values of C were used. This is because it is important for the model to not miss-classify cases of mind wandering.

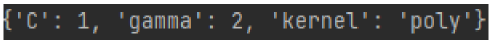
The scoring metric used to determine which parameters to keep was F1-Score(Macro).

This tuning was done with both the pooled and the non-pooled features, both with similar results which are shared below.

Table 4: Results of the tuned SVM on both pooled and non-pooled data from the test set

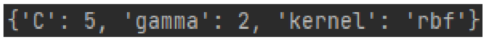
Classifier	Accuracy	F1-Score(macro)	MCC
Weighted SVM Pooling	0.8665	0.4642	-0.039
Weighted SVM Non-Pooling	0.8150	0.4490	-0.0488

The parameters chosen through the tuning process were as shown in the images below:



```
{'C': 1, 'gamma': 2, 'kernel': 'poly'}
```

Figure 4: The figure shows the values for the kernel, C value, and gamma value chosen through the tuning process using the non-pooled data.



```
{'C': 5, 'gamma': 2, 'kernel': 'rbf'}
```

Figure 5: The figure shows the values for the kernel, C value, and gamma value chosen through the tuning process using the pooled data.

The final results even after tuning are quite low. Although the classification models started to identify more cases as mind wandering, they were not correct.

Regardless, the results are more realistic than pre-tuning, where almost every data point was identified as non mind wandering.

4 Responsible Research

The applications of automatic detection of mind wandering are ever-increasing as almost every daily task is related to screens in one way or another. From detecting mind wandering during driving to cases during an online class or lecture, it can be used to help people maintain focus on their tasks to increase efficiency, or in the case of driving, reduce accidents.

However, even though the application of these systems is to increase the effectiveness in daily actions for people, there is an ethical concern stemming from the nature of the data that is needed for these systems to work, video data.

In this study, the video data used was collected from a website where participants agree to do certain tasks, in this case for the Mementos data set, it was to watch a couple of music videos and focus on them for their entire duration.

The participants were aware of the potential uses of the multimodal data collected and agreed with them. There was also additional data other than video and audio data collected which was not used for this study. The methods for collecting this data were approved by the university's Human Research Ethics Committee.

When handling the data for annotation, an offline application was used to ensure that the data being annotated is not in danger of being intercepted and leaked by any malicious actors.

Other than that, the data was used and handled on one device locally, without any public online repositories. The identities of the participants were not available and the faces were not used in any way for recognition, only their actions were used for mind wandering detection.

To reproduce the methods used in this study, access to the Mementos data set needs to be provided. This would mean whoever is interested should contact the university for more information.

5 Future Work

There are many suggestions and possible methods to experiment with for the future. This study was originally meant to train a residual network to do the classification and experiment with the architecture of the network to observe differences in performance and efficiency with changing final layers and methods of input. But due to lack of time, many of the experiments were abandoned and the scope of the study was lowered.

5.1 Training ResNet-18 or other networks for classification

Due to time limitations, there was no training done for the ResNet-18 model that was used. The model was used pre-trained on the ImageNet database, but adding additional training on top of that with the data from the Mementos dataset could provide more discriminative features for classification.

These features could provide substantial help to whatever model is then used for classification considering the current results.

Besides ResNet-18, other networks could also be used to study the use of learned features from neural networks in the automatic detection of mind wandering.

One architecture than can be used is the one shown in *Feichtenhofer et al., 2016*. Considering the results of the paper, it would be a good increase in complexity from a base ResNet-18 network and would take into account the temporal information.

Another suggestion that can be used is LSTM. Usage of this network can also allow for the study of a different representation of the temporal information from the data turning it into time-series data.

To add to the idea of turning the problem into a multivariate time series classification problem, ResNet has been shown very effective in these scenarios, as demonstrated by *Ismail Fawaz et al., 2019*, therefore looking into transforming the data could be beneficial for handling the temporal information and training networks.

5.2 Experimenting with different input pre-processing

Seeing that pre-processing of the input is a major part of the performance of a model, it is important that the correct pre-processing is done for the task that the model is handling.

Once again, due to the time limitations, it was difficult to effectively check the effects different pre-processing methods would have on the results, as the processing of extracting new features and re-fitting the classification models would take too long.

The pre-processing for this study was kept to be the same as the pre-processing used for ImageNet, but that pre-processing may not be the most fitting for this data set, although it is commonly used in papers that use pre-trained models that are trained on ImageNet.

With more time and resources, different pre-processing methods could provide much more clarity to more important parts of the original images, hence the extracted features would be more discriminative for classification.

5.3 Data augmentation or data collection

The data imbalance problem is a major reason for performance loss in machine learning. Because mind wandering data sets are mostly imbalanced, and the data extracted from the Mementos dataset is no exception, having a bigger amount of data would create possibilities for future studies to be more ambitious and use more complex systems for detection, such as very deep neural networks. These networks could then provide much higher performance, leading to more practical applications of mind wandering.

This data could be obtained by using data augmentation techniques that are fitting for the existing data set and video action recognition, or by just expanding the data set with more data.

Some data augmentation techniques that could be fitting for this data set would be techniques including the application of salt and pepper, gaussian blur, addition, or multiplication to simulate more of the possible environments of the ”In-the-wild” setup.

6 Conclusions and Discussion

In this paper, a study was presented on using features generated by a residual network to automatically detect mind wandering through supervised learning models as classifiers.

A definition of mind wandering that is appropriate for the data set and agrees with popular literature was discussed.

Due to time constraints, there was no experimenting done with pre-processing operations.

Several models were compared based on effectiveness and how appropriate they were for the data used, and Weighted SVM was chosen.

After tuning the parameters for the SVM, results showed that it was not able to perform better than a majority class classifier.

This performance was due to issues with data imbalance, features that were not discriminative enough, and a lack of pre-processing operations that were specific to the data. A major factor for the lack of experimentation was the time constraint.

Regardless of these results, neural networks and the learned features that can be extracted from them have been shown to perform really well in Computer Vision. Because of this performance shown in other Computer Vision tasks, through further study and experimentation with the methods suggested in section 5, this study can be expanded upon to achieve better results and answer all the research questions presented positively.

The core contribution of this study is comprised of the results of the performance analysis of several supervised learning models, mind wandering analysis, and data annotation for the continuation of this study.

References

Blagus, Rok and Lara Lusa (Mar. 2013). “SMOTE for High-Dimensional Class-Imbalanced Data”. In: p. 106.

Dudzik, Bernd et al. (2021). *”Collecting Mementos: A Multimodal Dataset for Context-Sensitive Modeling of Affect and Memory Processing in Responses to Videos”.*

Feichtenhofer, Christoph, Axel Pinz, and Richard P. Wildes (2016). *”Spatiotemporal Residual Networks for Video Action Recognition”.*

He, Kaiming et al. (June 2016). “Deep Residual Learning for Image Recognition”. In: pp. 770–778.

Hosseini, Seyedroohollah and Xuan Guo (2019). *”Deep Convolutional Neural Network for Automated Detection of Mind Wandering using EEG Signals”.*

Ismail Fawaz, Hassan et al. (July 2019). “Deep learning for time series classification: a review”. In.

Schooler, Jonathan et al. (2014). *The Middle Way: Finding the Balance between Mindfulness and Mind-Wandering.*

Singha, Subroto (2021). *”Gaze based mind wandering detection using deep learning”.*

Smallwood, Jonathan and Jonathan Schooler (Dec. 2006). “The Restless Mind”. In: pp. 946–58.

Steward, Angela et al. (2017). *”Face Forward: Detecting Mind Wandering from Video During Narrative Film Comprehension”.*

Van Gool, Luc (June 2008). “Action Snippets: How many frames does human action recognition require?” In.

Voulodimos, Athanasios et al. (Feb. 2018). “Deep Learning for Computer Vision: A Brief Review”. In: pp. 1–13.

Zhao, Yue, Christoph Lofi, and Claudia Hauff (2017). *”Scalable Mind-Wandering Detection for MOOCs: A Webcam-Based Approach”*.