USING NEURAL NETWORKS FOR FACIAL SENTIMENT ANALYSIS

By

Aaron Ward

Supervisor(s): Stephen Sheridan

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
B.SC IN COMPUTING
AT
INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN
DUBLIN, IRELAND
2018

Declaration

I herby certify that this material, which I now submit for assessment on the programme of study leading to the award of **B.Sc in Computing** in the Institute of Technology Blanchardstown, is entirely my own work except where otherwise stated, and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfillment of the requirements of that stated above.

Dated: 2018

Author: Aaron Ward

Abstract

Acknowledgements

/********* Insert acknowledgment here *******/

Table of Contents

List of Tables

List of Figures

Abbreviations

ANN	Artificial Neural Network
CNN	Convultional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
ReLU	Rectified Linear Unit
ML	Machine Learning
NLP	Neuro-Linguistic Programming

Chapter 1

Introduction and Background

The field of artificial intelligence and machine learning has become exponentially prominent in our daily lives. From business to social media, machine learning algorithms are being used to change the definition of efficiency and user experience. For example, Artificial intelligence algorithms are utilised by many large companies optimize the experience with image and voice recognition and photo searching (Deshpande, 2016). Also, Motor companies such as Tesla Motors use computer vision for their self-driving cars, which is a form of artificial intelligence.

The researcher and founded of convolutional neural networks (CNN's), Yann LeCun, became the director of Facebook's Artificial Intelligence department in 2013, and it is said believed that Facebook uses CNN's for it's facial recognition, user classification and tagging features (Deshpande, 2016). CNN's, in conjunction with recurrent nets, are also used for Facebooks DeepText feature. DeepText is a deep learning text-understanding engine used to comprehend and classify human generated textual content in over 20 languages (Abdulkader et al., 2017). In recent year, neural networks have been used in the field of medicine to better predict diagnoses and detection of cancerous tumours. For example, CNN's have been used by researchers for brain tumor segmentation (Havaei et al., 2015). The use of CNN's proved to be an appropriate method for tumour segmentation as the results can be given from a range of 25 seconds to 3 minutes (Havaei et al., 2015). Artificial neural networks have been used by radiologists for Computer-Aided detections systems (CADe) and Computer-aided Diagnosis

systems (CADx) to improve the accuracy of diagnoses, early detections and to minimize the time spent on evaluation by doctors (Firmino et al., 2014).

Many businesses depend on artificial neural networks for their business model as they can be applied to many industries and disciplines. According to Bhargava and Gupta (2017) artificial neural networks are used in a range of business applications such as forecasting of sales, classification of spending patterns, market targeting, risk analysis and bankruptcy prediction, to name a few.

Chapter 2

Literature Review

2.1 Analyzing and Detecting Employee's Emotion for Amelioration of Organizations

Subhashini and Niveditha make the opening statement that emotions usually do not any place in a work environment in current society. Although the expression of feelings is suppressed in places of work, they suggest that emotions can affect five major areas in competitive advantage. The five given aspects of competitive advantage are as follows: Intellectual Capital, Customer Service, Organizational Reactivity, Production, Employee appeal and retentivity (Subhashini and Niveditha, 2015). In order to counter this apprehensiveness to expression of emotions in the work place, Subhashini and Niveditha suggest the concept of a facial emotion tracking system that will map the facial expressions of an employees face as they enter the organization. Two related works are also given; Emotion Detections Based on Text and Emotion Recognition Based on Brain-Computer Interface Systems.

The system architecture given by Subhashini and Niveditha briefly describes the program. Most employees entering a building to an organization must swipe a card to clock into the work hours. They suggest that they have designed a new system that removes the need for card swiping but also implements emotion detection. The employee must look into a camera that will prove their presence in the vicinity but will also perform some emotion detection. The system was implemented in the C Sharp programming language and uses

skin tone segmentation to detect. The binary image is then converted to an RGB image and an inspection of every individual pixel if performed. If the RGB value is greater than 110, then the pixel colour is refactored to be a white pixel, otherwise it becomes a black pixel. This is done to make it easy to detect facial features in the video stream (Subhashini and Niveditha, 2015). Once detected, the image around the face is cropped. They then apply a Bezier Curve to the regions around the lips and eyes of the person being analyzed. The results of the person;s identity and emotional status are then stored within a database. They conclude their paper by explaining that this system can be used by management to gain an understanding of their employee's sentimental state.

2.2 Deep Learning for Video Classication and Captioning

Wu et al.'s paper provides an in-depth analysis on the methods for video classification and video captioning in terms of deep learning. They claim that because of the exponential growth in internet bandwidth and computing power, video communications are becoming more and more prevalent, therefore paving the way for new video understanding applications (Wu et al., 2016). They make reference to current implementations to prove the growth of interest in the field of computer vision and video analysis, notably the ImageNet challenge.

Move over, they go on to give brief description of the two "deep learning modules" that have been used for visual analysis: Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). It is explained that LeCun et al. that developed LeNet-5 made a break through when they developed a CNN using the Back-Propagation Algorithm. But its noted that this is limited in performance when the complexity of the tasks is increased. Deep belief networks were developed to train networks in a unsupervised manner in order to counter this problem (Wu et al., 2016). AlexNet, a CNN proposed by Krizhevsky et al. in 2012, introduced two way to increase the performance of CNN's using ReLU (Rectified Linear Units) and Dropout to decrease overfitting (Wu et al., 2016). Secondly, RNN's are

brought forward. Wu et al. explain the difference between CNN's and RNN's, stating that CNN's are all feed forward networks that do not use cycling, which can prove be disadvantageous when working with sequence labeling. Two issues can occur with RNN's: Vanishing Gradient and Exploding Gradients as short term memory is used when cycling through the network. The solution given to this is an RNN variant called Long Short-Term Memory (LSTM).

2.2.1 Image-Based Video Classication using CNN's and LSTM's

They state that Karparthy et al. researched the common architectures for learning spatial-temporal clues in large video datasets. It appeared that models using single frames as input achieve similar results as models using stacks of frames. From this, Simonyan and Zisserman proposed the idea of the Two-Stream Approach, because of the cost effectiveness and time consumption that come with training 3D CNN's. This Two-Stream approach involves training the CNN on single and stack frames concurrently. Both outputs are put through a score fusion. The result is the weighted sum of both scores (Wu et al., 2016).

Although Two-Stream is a good approach, it is not sufficient as it is not capable of dealing with long video clips. Therefore, LSTM's are utilized as they do not suffer from the problem of vanishing gradients. It has been found that CNN's and LSTM's compliment each other when working in conjunction with eachother (Wu et al., 2016). They conclude their paper with a summary of the written topics about, regarding the growth for the need for video understanding applications, the used of CNN's and RNN's, in addition to the variants of these deep learning modules.

2.3 Subject independent facial expression recognition with robust face detection using a convolutional neural network

As stated by Matsugu et al., difficulties may arise with facial recognition. In terms of a face being in a smiling-like state, could have different implications. As well as this, a facial recognition system should be able to work with a wide range of variability of faces. They then give some examples of past implementations such as facial recognition with rigid head movement by Black and Yackoob in 1995, and speak about how this does not meet the requirements of dealing with wide variance. A rule based system is proposed (Matsugu et al., 2003). With their model, layer trains on a module-by-module (module being the nose, eyes, mouth etc) basis. Meaning each layer trains on a certain facial feature. Each of the neurons perform an averaging of some local receptive fields then they use a skin tone detector to detect each module on the face (Matsugu et al., 2003). For training, Layer one and layer two are trained for 8 modules using back propagation. Layer three and four train on more complex feature detectors such as the mouth and eyes. The output is then sent to the rule based algorithm for handling variability and robustness.

The rule based algorithm takes the output of the CNN and measures the distance between the features. From these calculation, the rules are applied to determine if the person is in a laughing or smiling state. The rules are summarised as follows (Matsugu et al., 2003):

- The distance between eyes and lip get shorter.
- The horizontal length of lip gets longer.
- The eyes wrinkle.
- The gradient of lip from the end point to the end point increases.
- Detection of teeth increases.
- The edges (wrinkles) of cheek increase.

In conclusion, they received a 97.6 percent accuracy for 10 test subjects with 5600 images. They assert that their model is significantly more efficient as they only require one CNN due to their rule based algorithm, in contrast to Fasels implementation that uses two CNN's working in synergy with one another.

2.4 Neuromarketing - The Art and Science of Marketing and Neurosciences Enabled by IoT Technologies

A paper by Arthmann and Li describes the growing field of Neuromarketing with an opening statement: Advertisers recognise that there is a relationship between stimulating the emotions of a customer and influencing their actions. Online shopping has drastically affected the store sales, and it is also explained that more than 3500 stores have shut down due to bankruptcy in 2017 (Arthmann and Li, 2017). Their answer to this change in buying is Neuro Linguistic Programming. Neuro Linguistic Programming (NLP), not to be confused with natural language processing, is a form of observing the verbal and non-verbal communication of humans. Eye accessing cues (eye movements) are said to be linked to certain emotions or thoughts. Neuromarketing incorporates NLP and IoT devices to understand the consumer sentiment more extensively. Neuromarketing aims to remove marketing biases by utilizing the consumers subconscious. One example given by Arthmann and Li is the notion of facial coding and motion tracking, which is put in place to determine why consumer make certain decisions. Furthermore, they make their belief clear of retailers benefiting from this when it is put forward that artificial intelligence and machine learning will evolve, giving better results of consumer preference shifts. Additionally, These neuromarketing systems can replace thermoimaging people counting devices that clock people walking into stores that may not be eligible customers (children). Further examples are provided for these technologies. The describe a scenario when a customer is given an image of a product and a price in front of a webcam, and by performing facial coding and sentiment analysis, we may get a better sense of what the consumer is feeling. Arthmann and Li conclude their paper by

declaring the future potential of these systems using when integrated with ”always on” IoT technologies.

2.5 Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order

Lopes et al. open their paper by explaining the definition of facial expression. They describe it as facial changes that occur with someone due to emotional state or social communication. In terms of facial expression recognition software, a lot of systems give misleading accuracy results due to the overlapping of training and test data. They explain that some problems may occur. For example: dealing with ethnicity and variability of faces. Their response to this is training with one data set and testing with another to provide more accurate results (Lopes et al., 2017). They make reference to Liu et al's work on facial recognition by describing the three stages of training: Feature learning, which is responsible for extraction of all facial features. Secondly, Feature Selection, that selects the best features to learn. And lastly, the classifier, that each expression has one specifically allocated to. Lopes et al. then move on to explain convolutional neural networks (CNN) on a high level. Firstly, the CNN is comprised of a convolution layer, that if given a kernel size. This kernel shifts along the image given to generate a map. This is followed by sub-sampling. Sub-sampling is used to reduce the map size to increase the accuracy in variance. Lastly, the fully-connected layer is introduced. This is a neural networks that has fully connected neurons to it's previous layer (Lopes et al., 2017) and it is stated that CNN's using supervised learning use gradient descent. They explaing that of all the facial recognition methods, CNN's prove to be very advantageous because they can use raw image data input for an accurate prediction.

A related work section is then brought forward to explain similar models that have been developed. Much progress has been made in the field of neural networks have come about in recent years. This is due to advances in GPU technologies and computing power Lopes et al.

(2017). One example of relevant work they provide is the work done by Song et al., in which a CNN was developed for a mobile phone application for facial expression recognition. This CNN used image augmentation techniques, due to the lack of public data for their network, to prevent over-fitting. This increases the amount of data available for training the network (Lopes et al., 2017). Song et al. received and accuracy of 99.2 percent using the CK+ dataset. A CNN with 15 layers was developed by Burkert et al that achieved similar result, of 99.6 percent. They point out that although this network achieved high results, they may prove to be misleading as is it not made clear that their training and test datasets were different (Lopes et al., 2017). They then go on to compare the related works by expressing the flaws of using over-lapping data and the lack of emotion expressions classified.

2.5.1 Facial Expression Recognition System

Following the introduction and related work, Lopes et al. provide a prerequisite understanding for their model. Their model has two stages: the training stage and the testing stage. The training stage consists of a few preliminary steps. Firstly, new images are made from existing ones in the dataset to increase the trainable data. This is done using a method proposed by Simard et al. called ”synthetic sample generation”, which involves rotating and skewing the existing images. For every photograph that exist, and additional 70 are made, adding noise to the data. This synthetic data is only used in the training stage and advantageous as it allows the model to handle variance in an image (Lopes et al., 2017). Secondly, to address the problem in alignment with of facial features, the notion of rotation correction is introduced that aligned the images using the eyes as the horizontal axis. Image cropping is then used to reduce the amount of background noise as it is said to decrease the accuracy and overall performance of the CNN. This is done by detecting only features that are valid of expression classification and croping th image around them, excluding the neck, ears and background from the image. Down sampling is then applied for reducing the size of the image, making it 32 x 32 pixels in size. Brightness and contrast can cause problems with images, therefore intensity normalisation is applied to lower these aspects of the image (Lopes et al., 2017).

As for the testing stage, the same methods are used as the training stage. The CNN outputs the predicted emotional expression with the following number ID's:

0 - Angry

1 - Disgust

2 - Fear

3 - Happy

4 - Sad

5 - Surprise

2.6 Relevant Work and Critical Analysis

Subhashini and Niveditha, 2015's work on analyzing and detecting employee's emotions for organizations ties in very well with the proposed project as it involves using sentiment analysis on subject to gain an underlining understanding of their emotions that may or may not be expressed verbally. Their approach to skin tone segmentation for detecting a humans face may prove very beneficial to the proposed project. The main research goal was to achieve employee identification in conjunction with facial emotional analysis and they were successful in execution, They used a Bezier Curve on the subjects lips and eye to detect the emotion expressed by analysing the gradient of the curve. However, Although this paper proved that the project was a success, there are some inconsistencies. For example: in the related work, they do not explain how the work is related and only give titles. Secondly, the results show no code snippets or pseudocode to explain the implementation of the system. Only screen shots of the user interface are provided. Furthermore, there are more absences of proof. The paper is lacking statistics and graphs to display the accuracy or progress of the systems performance and there are some bold statements used that are not back up by citations like when it is said that "emotions were considered a forbidden topic in the working place". Despite these weak areas in the paper, a good aspect of this system is the use of real

life subjects used in testing.

The review of Deep Learning for Video Classification and Captioning by Wu et al. provides an in depth look into the aspects of different neural networks and what are they strong and weak points. The motivation for their research is driven by their claim that video communications is growing and that their needs to be better applications for video understanding. The relevancy of this paper provides the concept of the ”Two-stream” architecture. Although it is not planned to develop two convolutional neural networks (CNN), is it sought after to develop a score fusion algorithm, similar to the one mentioned in this paper. A CNN will be developed and the application will utilize a tone analyser for voice sentiment analysis and the two scores will be combined by such an algorithm to provide the weighted sum of the two scores. As this paper is very in depths and draws good comparisons between the different techniques that can be used for video classification. Despite the quality of this paper, some aspects need improvement. There is heavy usage of words like ”we” used. Also, some statements are made by (Wu et al., 2016) that are not cited to supported their claim. This is evident when its said ”As deep learning for video analysis is an emerging and vibrant field...” (According to whom?).

Subject independent facial expression recognition with robust face detection using a convolutional neural network by Matsugu et al. illustrates the difficulties that may arise when performing facial recognition. They highlight the problems that may occur in terms of being able to handle variability of subject faces, and certain angles. Their approach to this problem is addressed by implementing a rule based algorithm that analysis the results given from the CNN. Furthermore, their model is designed to be segment and be trained on specific facial features instead the face as a whole. Their model proves to be a success as they score an accuracy of 97.6 percent, also they do not require a second CNN working concurrently to achieve similar results as other models have done so previous, which can be cost effective. Some similarities arise between this paper and the proposed project, they both use sentiment analysis and require the ability to handle a wide range of variability. This proves beneficial

to the proposed project as it provides inspiration to used a rule based algorithm for determining emotions. Even though this paper is well written, there are some issues. In certain parts there are abbreviations to words given without the full word be given prior which can cause confusion to the reader. For example: ”FP neurons”. In addition to this, their model is specific for smiling faces and doesn't accommodate for other emotions, which should be at least provided in a further work section.

Arthmann and Li's paper titled Neuromarketing The Art and Science of Marketing and Neurosciences Enabled by IoT Technologies is a promising insight to the field of neuromarketing. They recognize the association of online shopping and loss of sale for retail stores, and give example of how neuro-linguistic programming and IoT technologies can be used as a combination to understand their customers sentiment. Furthermore, it's heavily argued that the integration of AI and machine learning will evolve this concept to understand the thoughts of consumers and further tackle loss in productivity. This proves relevant to the proposed project as it is very similar. They both aim to gain a deeper understanding of human sentiment that may not be express verbally. Although this is possibly the most interesting paper of this review, it is severally lacking citations. Also, there are a lot of assumption brought forward with no clear indication as to how this knowledge is known. Additionally, it is also assumed that people will adopt these ”always on” IoT devices and agree for their physical aspects to be used for consumer targeted marketing.

2.7 Conclusion

In conclusion, the five papers reviewed topics in relation to convolutional neural networks, facial sentiment analysis, emotion detection and current applications in the real world. Relevant aspects include detecting employees emotions in a work environment, implementing a score fusion algorithm for achieve a summation of two sentiment detecting technologies, the idea of segmenting facial features while training to accommodate for variability and the notion using human emotion understanding for a business solution.

Chapter 3

Methodology

3.1 section header 1

3.1.1 Subsection header 1

Chapter 4

System Design

4.1 section header 1

4.1.1 Subsection header 1

Chapter 5

Implementation

5.1 section header 1

5.1.1 Subsection header 1

Chapter 6

Testing and Evaluation

6.1 section header 1

6.1.1 Subsection header 1

Chapter 7

Discussion

7.1 section header 1

7.1.1 Subsection header 1

Chapter 8

Conclusion and Further Work

8.1 section header 1

8.1.1 Subsection header 1

Bibliography

Ahmad Abdulkader, Aparna Lakshmiratan, and Joy Zhang. Introducing deeptext: Facebook's text understanding engine, 2017. URL https://code.facebook.com/posts/181565595577955/introducing-deeptext-facebook-s-text-understanding-engine/.

Christopher Arthmann and I-Ping Li. Neuromarketing the art and science of marketing and neurosciences enabled by iot technologies, 2017. URL https://www.iiconsortium.org/pdf/2017_JoI_Neuromarketing_IoT_Technologies.pdf.

Nikhil Bhargava and Manik Gupta. Application of artificial neural networks in business applications. *GEOCITIES*, 1:3–4, 2017.

Adit Deshpande. A beginner's guide to understanding convolutional neural networks. https://adeshpande3.github.io/adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/, 2016. [20th July 2016].

Macedo Firmino, Antnio H Morais, Roberto M Mendoa, Marcel R Dantas, Helio R Hekis, and Ricardo Valentim. Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects, Apr 2014. URL https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-13-41.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron C. Courville,

Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *CoRR*, abs/1505.03540, 2015. URL http://arxiv.org/abs/1505.03540.

Andr Teixeira Lopes, Edilson de Aguiar, Alberto F. De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 61(Supplement C):610 – 628, 2017. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2016.07.026. URL http://www.sciencedirect.com/science/article/pii/S0031320316301753.

Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5):555 – 559, 2003. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(03)00115-1. URL http://www.sciencedirect.com/science/article/pii/S0893608003001151. Advances in Neural Networks Research: IJCNN '03.

R. Subhashini and P.R. Niveditha. Analyzing and detecting employee's emotion for amelioration of organizations. *Procedia Computer Science*, 48(Supplement C):530 – 536, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.04.131. URL http://www.sciencedirect.com/science/article/pii/S1877050915006407. International Conference on Computer, Communication and Convergence (ICCC 2015).

Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. Deep learning for video classification and captioning. *CoRR*, abs/1609.06782, 2016. URL http://arxiv.org/abs/1609.06782.

Appendices

Appendix A
sample section header