

PhD Study #3: In Defence of ML/CNN for the SensorAble Research Project

David Ruttenberg^{1*}, Kaśka Porayksa-Pomsta², Sarah White³, Joni Holmes⁴

^{1, 2} University College London, Institute of Education, Culture, Communication and Media Department, University College London Knowledge Lab

³ University College London, Institute of Cognitive Neuroscience, Development Diversity Lab

⁴ University of Cambridge, MRC Cognition and Brain Science Unit

* Correspondence: david.ruttenberg.18@ucl.ac.uk

Abstract

Animals and humans use a midbrain structure to coordinate and process relevant visual and auditory stimuli while suppressing distracting information. In modelling this assembly and managing both environmental and physiological stimuli using *engineering principles*, my research aspires to deep learning models that sense, categorize and alert autistic individuals of ecological distractions, biophysical cues and other multimodal input that—left unchecked—could decrease individual focus and increase distractibility and anxiety. The designs that follow are based upon valid and reliable constructs presented in recent, peripherally related research, including: (i) a framework for developing adaptive intelligent user interfaces that enhances user experience (Johnston et al., 2019); and, (ii) convolution neural networks (CNNs) that improve expression recognition through emotion-modulated attention (Barros et al., 2017). My intention is to weave a compelling and explicit rationale as to how and why deep learning models make the most sense when learning tasks derived from image, time-series and text-data and applying these to the SensorAble Research Project.

Keywords: Autism Spectrum Condition, Attention, Focus, Machine Language, Convolutional Neural Networks, Cross-Channel Convolutional Neural Networks, Distractibility, Anxiety, Focus, Adaptive Wearable.

1. Introduction

Previous research proposes neurocomputational models that learn emotional expressions and subsequently modulates emotional recognition (Barros et al., 2017). Identifying emotional expressions in a cluttered environment is implemented through input images sensed and processed by a convolutional neural network. Experiments have shown that with CNN, attention improves recognition when used as a cognitive modulator. By re-engineering these neural processes, and by substituting different input(s), I propose a nearly identical deep learning model that may improve not just emotional-modulated attention, but *overall* attention that aims to filter stimuli and subsequently alert autistic individuals to myriad and heterogenous cues that causes distraction, anxiety and other undesirable conditions.

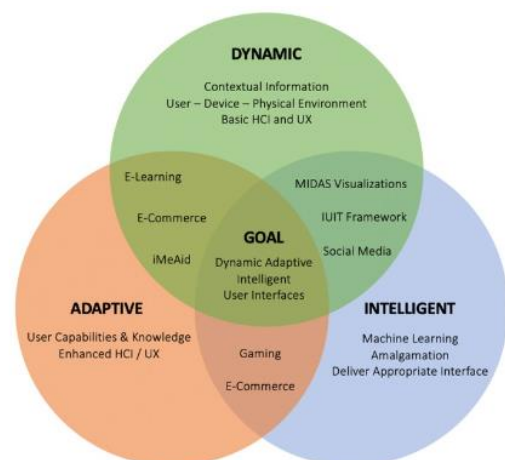


Figure 1: Engineering Components of Adaptive Systems

At a 35,000' level, three engineering components are required to deliver an adaptive system that enhances

user experience (e.g. greater focus, reduced distraction and attenuated anxiety). These appear in Figure 1 and

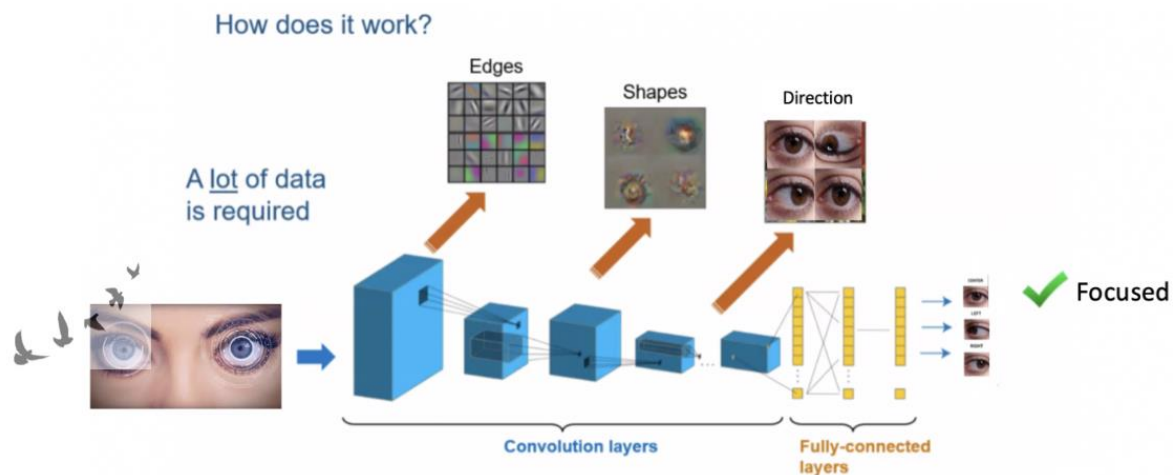


Figure 2: How does SensorAble aim to recognise distracting stimuli?

include *dynamic, adaptive and intelligent interfaces*—of which—research has identified an intersectional gap in their interactivity (Johnston et al., 2019). Specifically, *dynamic* systems understand the user, their device and their environment to provision a *basic user experience*. *Adaptive* interactivity extends these basics by including a user's capabilities, accessibility issues and *flow within the ecological/physiological space*. The *intelligent* element employs machine learning algorithms that deliver a *personalised user experience* based on her/his unique objective(s) and/or thresholds. Ultimately, this decreases the “user's cognitive load and enhances their cognitive and interactive experience” (Ibid, 2019, 32).

1.1. Redefining the research questions:

Manipulating previously identified neurocognitive, psychological and social science research questions is necessary—now from a re-imagined engineering perspective, specifically:

Q1.: Can ecological and physiological data help train deep learning (DL) models to distinguish distracting stimuli and localize their position?

Q2.: Can a trained model be used when comparing attentional thresholds and create subsequent data alerting a user when their distractibility and anxiety increases (e.g. based upon visual, sonic, inertial and physiological inputs)?

Q 3.: Can DL models be used to identify and predict disrupting stimuli before they occur?

Q 4.: Can these models be trained to overcome the paradox of typically narrowing results versus the ever-widening heterogeneity in autism?

Q 5.: Can this paradox be generalized to a wider autistic community beyond the laboratory; and, can DL models be operationalized and correlated to field-based distractions bound by *in situ* accounts?

1.2. Blending engineering with biological neural mechanisms

Processing visual spatial, auditory attention and physiological stimuli occurs within several areas of the brain. One such area—the superior colliculus (SC)—is the primary neural mechanism integrating stimuli while processing target selection and motor consequences (Driver, 2001; Krauzlis, 2013). From an engineering perspective, neurophysiologists have studied SC and engineered how it assimilates multimodal impulses by means of perceptual cue computations that trigger attention (Ursino et al., 2014; Bauer et al., 2015).

United among these discoveries is the proposition that selective attention is modulated by the affective implication of sensory inputs (Vuilleumier, 2005). Behaviourists have shown that attention to expressive cues (rather than neutral prompts) are reflexive and involuntary; that is, visual targets expressing an emotional stimulus are recognized more quickly than those without analogous emotional indicators (Eastwood et al., 2001; Williams et al., 1996).

Blended training (resulting from assorted input signals) increases accuracy, suggesting that engineering of multimodal sensory inputs may provision more robust integration when distractibility is present (Castellan et al., 2008). Predictably, the preponderance of machine learning (ML) studies apply to *singular* modalities and reach acceptable performance levels for

just discrete tasks (Li et al., 2015; Viola et al., 2004). This represents a challenge as typical ML training (e.g. *face detection* or *singular domain stimuli*) does not support emotional recognition or, for that matter, SensorAble's aims.

Remember that SensorAble aspires to environmental and physiological inputs—not isolated facial images or body movements. Figure 2 provides a graphical example of a singular visual distracting stimuli affecting pupillary and/or gaze data—the latter of which is used to train a DL model. Transforming eye data streams to a time frequency image could provision ML.

Figure 3 shows how SensorAble proposes building on singular training by provisioning multimodal stimuli across audiometric, physiological/anxiety and inertia files to train CNN. Each data stream is converted from a fundamental state using time-frequency transformations that leverage greater validity and reproducibility during DL image training procedures.

SensorAble dimensions formerly successful training of expressive recognition and through CNN models that distinguish amid facial and body lexes. Those studies and their related engineering/designs may scaffold SensorAble's model from an emotional-basis to a perception-based system using *multimodal* sensory inputs (e.g. sonics, inertia, galvanic skin response (GSR), electrodermal response (EDR), electrodermal activity (EDA), skin conductance response (SCR)—and optics). This may offer significant advantages particularly in heterogenous conditions like autism by classifying hundreds of thousands of bio/ecological data points per mode and then labelling them as distracting.

From an engineering view, SensorAble utilises these multiple stimuli when replacing singular visual data to train and enable CNN to discriminate among distracting and focussing stimuli (conveyed by ecological and physiological activity in Fig. 3) rather than just facial expression or movement cues. In so doing, SensorAble may become an Intelligent User Interface (IUI) that functions in combination with Machine Learning (ML) algorithms to tailor a user experience that “provide[s] a better user journey” (Johnston et al., 2019, 33).

1.3. Framework definition

There are myriad challenges to this framework: “the main [one] being the *type* of ML algorithms that...work across more than one domain” (Johnston et al., 2019:34). As SensorAble seeks to train CNN on multimodal sensory data—including visual, sonic, inertia and physiological types—researchers discovered how algorithms assist an “adaptive layout of an interface...with user flow, usability and functionality aspects” (Ibid, 2019, 34). It is highly critical, then, to not

only inform and train systems of both meaningful and multiple eco/biological data, but of sensory inclinations, thresholds and distinctive potentials of *each user*.

1.4. Improving and extending an already working system

Scientists have shown that even when CNN input is composed of a single image sequence, “model training autonomously learns separate cue-specific filters” (Barros et al., 2017, 105). Unlike traditional systems that use labelling for learning methods, *probability distributions* (a more contemporary technique) enables image localization through selective attention. This is an important consideration as SensorAble may also be able to substitute alternative stimuli to triangulate spatial calculations (e.g. Figures 3 and 4 utilise three and two domains to elicit localization).

By combining multiple sensory inputs, stimuli localization has proven possible the harmonization of regions corresponding to directional and probabilistic estimates. SensorAble aspires to similar feed-forward methods that evaluate distracting stimuli against an individual's pre-defined thresholds. Comparably, scientists have calculated combinations of facial properties and body movements to detect emotionally relevant image areas. By *substituting* ecological and physiological data streams for these facial and body data, SensorAble proposes a re-engineering of similar combinations and calculations.

This study anticipates an analogous strategy using specific inputs comprised of: (i) single image sequences of eye tracking, pupillary and gaze fixation data; (ii) audiometric sensing of ambient amplitude, frequency and spatial distribution; (iii) inertial measurement units tracking individual head sway; and, (iv) physiological measures sequencing individual GSR, EDR, EDA and/or SCR. These inputs may provide critical training data across various domains, sensory inputs, classifications, filters and response triggers for heterogenous use (see Table 1). Further, SensorAble may rely on engineering and sensor data that is cloud-stored (e.g. Google BigQuery, AWS ML, etc.) aiding in relation to security and performance.

2. Deep attention model

The SensorAble model aspires to hierarchical learning and selective attention using CNN. This project differs from *traditional* CNN-based approaches in two aspects: first, SensorAble's input stimuli are composed of an *entire* physiological and ecological scene. By way of example, any combination including or eschewing people expressing emotions, anxiety or related bodily

data may or may not be included and combined with ecological data demonstrating distracting stimuli. Second, the network is trained to (a) *localize* where distraction(s) and/or anxiety(ies) exist; and, (b) *identify* if detected distractions or anxieties cross thresholds that

attract the model's attention along with any number of subsequent and related triggers (e.g. a haptic alert, audiometric filtering, coaching, etc.—listed as response triggers in the Table 1 below).

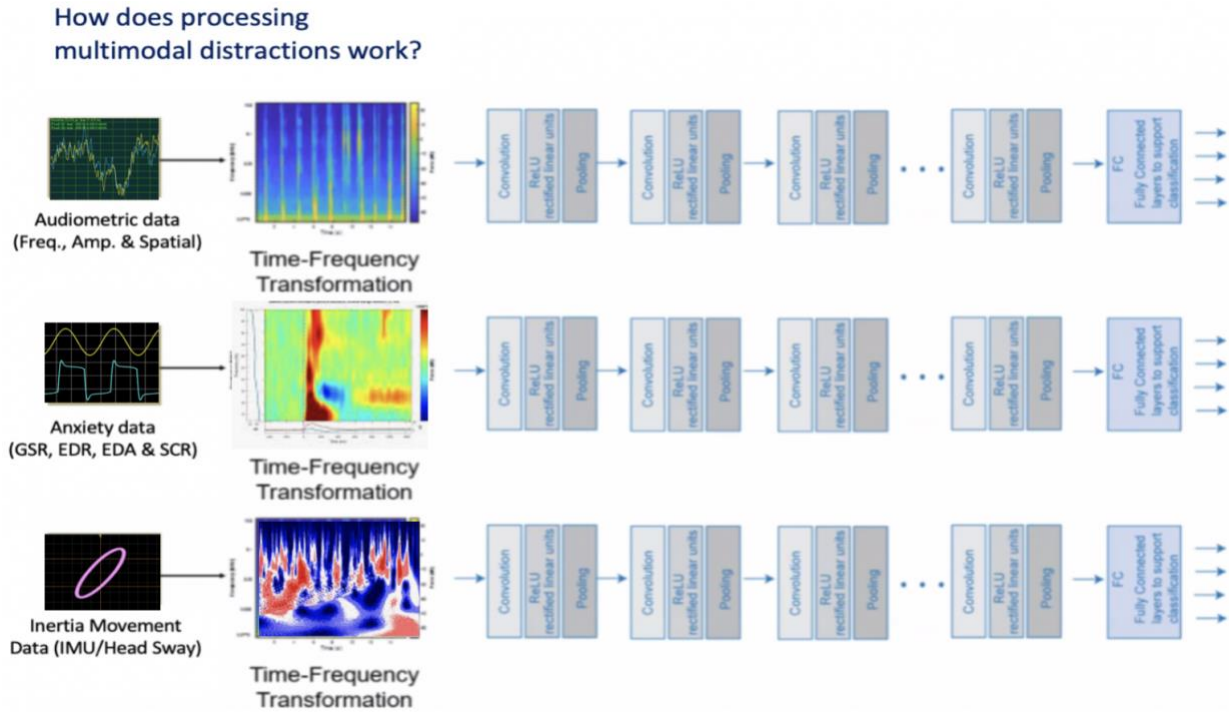


Figure 3: Combined multi-sensory inputs (across three domains) may provide greater accuracy when training models to localize stimuli and respond with subsequent alerts.

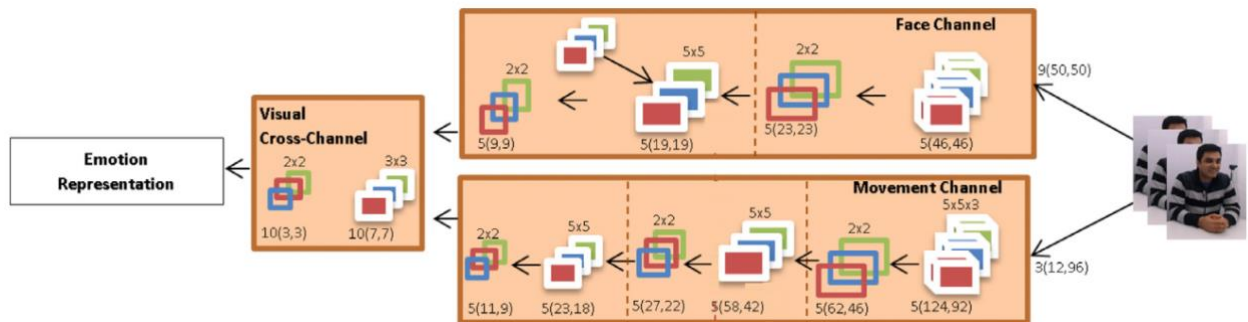


Figure 4: A two-domain training example utilising optical face and movement data to localize emotional stimuli.

While many contemporary CNNs learn hierarchical descriptors from input stimuli and then describe input data into smaller, highly abstract representations, SensorAble—like those studies from Barros, Johnston et al., (2017; 2019)—does not use a strict classification technique. Rather than learning tiered contours and shapes used in general image recognition tasks, convolutional units instead train on sensorial and spatial information sometimes *originally existing* as images and other times *transformed into* visual data (Speck et al., 2016).

Traditional CNNs train upon image libraries commonly assembled from digitized photographs, video/film frames, etc. SensorAble must use re-formed stimuli adapted into time-frequency transformations (e.g. audio or biometric signal turned into visual form). Explicitly, ecological audio stimuli are filtered first and later optimized from a frequency and amplitude perspective to better match human hearing spectrums and traits. This re-shaped bandwidth and amplitude affords supplementary benefits leading to more efficient training, processing, recognition and altering of resultant triggers due, in no trivial part, to the reduction size of initial data streams—much of which are outliers to human hearing and cognition.

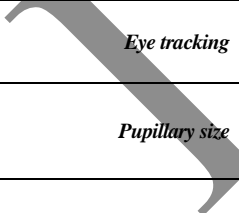
As mentioned previously, traditional models have successfully detected emotional events using facial expressions and body movements. Their associated convolutional units trained on feature processing from entire images. More recently, newer systems differ from

this type of classification task by not necessarily tuning convolutional units to designate forms. Rather, they identify *where* an expression is located *within* an image. This hierarchical localization represents a level of preciseness within a sub-region of an image. In these systems, a CNN's first layers learn how to detect Regions of Interest (ROI) that are then fine-tuned in successive, deeper layers. Because pooling units increase spatial invariance, they are applied to the last layers only, leaving earlier layers composed of only stacked convolutional units. Localization training such as this uses learning strategies based on probability density functions (Barros et al., 2017, 2015).

3. Filtering for perception and responsiveness

Barros et al.'s models were trained with full images and without dissection (e.g. facial expression and body movement were processed homogenously by convolutional filters). Once trained for combined visual stimuli, filters were then apportioned to either facial or movement reactivity. Using a deconvolution process in the network's last layer, the system reflects learning of differing modality detection. For example, filters 1–3 learned facial while filters 4–6 trained on movement.

SensorAble aspires to employ distraction onset and localization training to then alert and filter ecological

Domain	Sensory Input	CNN Classification	Response Triggers
Visual			
 Eye tracking	Infrared camera detection of pupil centres corneal reflections (PCCR). Reconstructs direct analogues of pupilometer measures from pupil diameter. Measures pupillary fixation, duration, saccade length and visit counts.	Classification of ~ 2,445,504 images from 1,474 persons yielding a model- and appearance-based (AB) classification and gaze estimation. The model is tested against existing AB CNN approaches, achieving better eye gaze accuracy with significantly fewer computational requirements.	V.T.1.: Sequencing patterns for attention, distraction and memory V.T.2.: Capable of measuring direction, intensity and emotion. V.T.3.: Works with V.T.1. for localization disturbances.
Pupillary size			
Gaze fixation			
Sonic			
Amplitude—perceived loudness or softness of sound	Quantified using multipattern transducers (microphones) in a binaural configuration and then converted to images using Time Frequency Transformation (TFT).	100M data set consisting of 70M training, 10M evaluation, 20M validation videos consisting of ~5.4M hours of audio. Each file is classified and identified from a set of ~31k labels.	S.T.1.: Adjusts loudness pending amplitude classification and user preference. S.T.2.: Adjusts equalisation based upon frequency classification. S.T.3.: Adjusts frequency spectrum pending sidechain input
Frequency—perceived pitch including timbre, resonance and Q			

<i>Spatial location—perceived location of sound in a stereophonic field</i>			of frequency modulated volume. S.T.4: Adjusts distracting sonic localization from actual spatial position to desired focal point.
Inertia			
<i>Head sway</i>	Quantified by IMU detection of vertical standard deviation of head position over a duration of one minute. The head face position is estimated from the detected eyes positions resulting from V.T.1.-3.	Human activity recognition (HAR) is a classification task for recognizing human movements. HAR uses inputs signals from videos or multichannel time-series, including inertial measurement unit (IMUs). Traditional HAR segments sequences using a sliding-window approach, extracting hand-crafted features from sequences, while training a classifier for assigning certain action sequence labels.	I.T.1: Works in tandem with visual and sonic domains to determine disturbance, lack of focus and localization.
Physiological			
<i>GSR/EDR—arousal, sweat, fear, anger, startled response</i>	Quantified by applying an electrical potential between two points of skin contact and measuring the resulting current flow between them.	Multiple heterogeneous sensors provide reliable data. Adaptive CNN models have improved emotion classification accuracy and reducing model instabilities. Effective spectrogram feature extraction and multimodal classifiers use two features as input at the first layer of a fully connected network.	P.T.1: Works in tandem with V.T.2. (pupillary size) to determine anxiety. P.T.2: Works with visual, sonic and I.T.1. (inertia) to determine onset of distractions and localization features.
<i>EDA—cognitive states, arousal, emotion and attention</i>			
<i>SCR – fear, anger, startled response, orienting response</i>			

Table 1: Domain, Input, CNN Classification and Response Triggers

and physiological data to the benefit of the individual. Once the system is successfully trained to filter for perception types, as previously mentioned, SensorAble aspires to leverage response triggers to clarify offensive and/or distracting stimuli (e.g. *spatially re-align* audio data that does not correlate to active focal points; *attenuate* unpleasant frequencies or amplitudes that may be inversely proportional and out-of-phase to disrupting signals; and/or, *predict* environmental cues that may exacerbate anxiety but are then modulated by haptic alerts prior to onset).

3.1. A one-sided attentional model

The two-stage hypothesis of emotional attention states that attention stimuli are first processed as a fast-forward signal by the amygdala complex, and then used as feedback for the visual cortex (Bullier, 2001). While theory provides for multiple connections between complex and cortex, previous CNN models like Barros et al.'s use one-sided modulation; that is, from *attention to perception* modelling. Based on theory, these models begin as fast-forward processing of attention and then use detected regions as perception inputs.

These systems use attentional features to modulate perception; specifically: (i) images are fed to attention CNN; (ii) a region is obtained; (iii) the model detects

face and/or body features; and, (iv) filters integrate in a second convolutional layer of a Cross-Channel Convolution Neural Network (CCCNN).

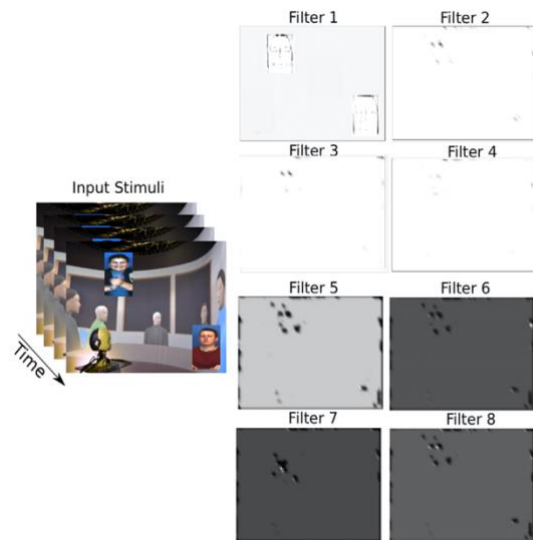


Figure 5: Last-layer filtering and whole feature extractions.

In turn the second layer of existing systems typically proceeds convolution of movement which occurs in a third layer to discern similar features (no longer

physiognomy but now *gesture*). This is depicted in Figure 6 below.

4. Experimental results

These existing models were evaluated using three different stimuli encompassing face, body movement and a combination of the two. Selective attention leveraged three scenarios including none, one and two displayed expressions. Results with one expression presented by a *single face* yielded top-20 accuracy of nearly 88%. With *one body* movement, accuracy was 85%. The best results were achieved by *combined cues* yielding >93%. Not surprisingly, two expressions/two persons presentations resulted in a drop of 76% for face and 68% for movement. When both expressions were present across two people, accuracy increased to ~85%.

5. Discussion

Regardless of models previously experimented or proposed, convolutional neural networks require immense data quantities to train properly. Barros et al.'s network has many connections to train and thus demands more computational effort because it is a CCCNN. However, once trained, computational costs for recognition demand are surprisingly low. The researchers reported that “a Core I5 computer with a graphic card Nvidia Quadro K620 [produced a typical] one *training* epoch of our CCCNN with attention [of] around 4.8 minutes, while one forward *pass* took 0.4 s” (Barros et al., 2017, 111).

6. Conclusion

Differing from traditional CNN-based classification tasks, SensorAble proposes a model with probability

distributions that indicate focus on attention. Similar to more recently structured systems, these models use unlabelled expressions to identify attention regions, being able to distinguish between emotional and neutral expressions that are substituted with visual, sonic, inertia and physiological data streams. Similar models have shown success as attention modulators for a Cross-Channel Convolution Neural Networks (CCCNN), improving upon CNN recognition capabilities. This is an applicable extension to the multitude of sensory modalities SensorAble aims to modulate with conjoined data input and compound response triggers.

Similar to CCCNN, SensorAble seeks to derive localization from an entire eco/physiological scene, where distracting stimuli/data are trained on the network which then subsequently learns where to focus. When expressions are observed, SensorAble concentrates on those stimuli that both *exceed a threshold* and *match a focus, distraction and anxiety profile* personalized by the individual. As in previous studies, SensorAble aspires to differentiate between two non-neutral stimuli, even if only one presents an attentional peak that may or may not correlate to an individual's threshold/profile. Like other CCCNN, this project aims to learn filtering algorithms that react to multiple stimuli even if the one or more may not have been explicitly defined.

From an engineering prospective, selective attention modulated by a stimulus (e.g. affective behaviour), provides evidence that salience influences responsiveness. These CCCNN use top-down mechanism simulating the selective, neuronal attention in the brain where distracting expressions attract more attention than neutral lexes.

Finally, Barros et al., suggest that “the integration of auditory information as a perceptual cue would give

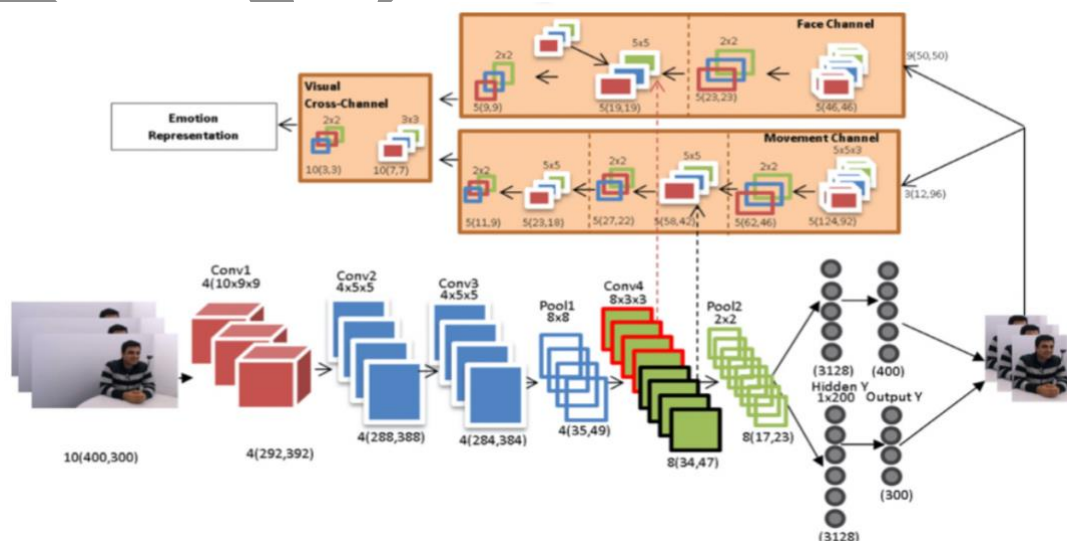


Figure 6: Flowchart of a Cross-Channel Convolutional Neural Network (CCCNN) processing facial and bodily movement data streams.

[their] model a more robust multimodal mechanism for emotional attention. Therefore, we expect that the addition of auditory information would increase precision...and approximate our computational approach to neurobiologically-motivated neural mechanisms for multimodal integration and attention” (Barros et al., 2017, 113).

7. Acknowledgements

The author wishes to acknowledge and greatly thank Professor Kaśka Porayska-Pomsta for her guidance and inspiration.

8. Reference

- Barros, P., Parisi, G. I., Weber, C., & Wermter, S. (2017). Emotion-modulated attention improves expression recognition: A deep learning model. *Neurocomputing*, 253, 104–114. <https://doi.org/10.1016/j.neucom.2017.01.096>
- Bauer, J., Magg, S., & Wermter, S. (2015). Attention modeled as information in learning multisensory integration, *Neural Netw.* 65 44–52.
- Bullier. (2001). Integrated model of visual processing, *Brain Res. Rev.* 36 (2), 96–107.
- Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech, in: C. Peter, R. Beale (Eds.), *Affect and Emotion in Human–Computer Interaction, Lecture Notes in Computer Science*, vol. 4868, Springer Berlin Heidelberg, pp. 92–103, doi:[10.1007/978-3-540-85099-1_8](https://doi.org/10.1007/978-3-540-85099-1_8).
- Chen, S., Tian, Y., Liu, Q., & Metaxas, D.N. (2013). Recognizing expressions from face and body gesture by temporal normalized motion and appearance features, *Image Vis. Comput.* 31 (2) 175–185. <http://dx.doi.org/10.1016/j.imavis.2012.06>.
- Driver, J. (2001). A selective review of selective attention research from the past century, *Br. J. Psychol.* 92 (1) 53–78.
- Eastwood, J.D., Smilek, D., & Merikle, P.M. (2001). Differential attentional guidance by unattended faces expressing positive and negative emotion, *Percept. Psychophys.* 63 (6) 1004–1013.
- Jin, Q., Li, C., Chen, S., & Wu, H. (2015). Speech emotion recognition with acoustic and lexical features, in: *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4749–4753, doi:10.1109/ICASSP.2015.7178872.
- Johnston, V., Black, M., Wallace, J., Mulvenna, M., & Bond, R. (2019). A Framework for the Development of a Dynamic Adaptive Intelligent User Interface to Enhance the User Experience. *Proceedings of the 31st European Conference on Cognitive Ergonomics - ECCE 2019*, 32–35. <https://doi.org/10.1145/3335082.3335125>
- Krauzlis, R.J., Lovejoy, L.P., & Zénon, A. (2013). Superior colliculus and visual spatial attention., *Ann. Rev. Neurosci.* 36 (1) 165–182, doi:10.1146/annurevneur-062012-170249.
- Li, H., Lin, Z., Shen, X., Brandt, J., & Hua, G. (2015). A convolutional neural network cascade for face detection, in: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5325–5334.
- Lin, Y.-L. & Wei, G. (2005). Speech emotion recognition based on HMM and SVM, in: *Proceedings of the 2005 International Conference on Machine Learning and Cybernetics*, vol. 8, IEEE, pp. 4898–4901.
- Speck, D., Barros, P. Weber, & C., Wermter, S. (2016). Ball localization for RoboCup soccer using convolutional neural networks, in: *Proceedings of the 2016 RoboCup International Symposium (RoboCup’2016)*.
- Ursino, M., Cuppini, C., & Magosso, E. (2014). Neurocomputational approaches to modelling multisensory integration in the brain: a review, *Neural Netw.* 60 141–165.
- Viola, P. & Jones, M.J. (2004). Robust real-time face detection, *Int. J. Comput. Vis.* 57 (2) 137–154.
- Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention, *Trends Cognit. Sci.* 9 (12) 585–594.
- Williams, J.M.G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology., *Psychol. Bull.* 120 (1)

DRAFT