Exploration of Design Space and Runtime Optimization for Affective Computing in Machine Learning Empowered Ultra-Low Power SoC

Yijie Wei, Kofi Otseidu, Jie Gu
Northwestern University, Evanston, IL
{yijiewei2019, kofiOtseidu2022}@u.northwestern.edu, jgu@northwestern.edu

ABSTRACT

The incorporation of artificial intelligence into the rapidly growing IoT devices demands a high level of built-in intelligence, e.g. machine learning capability at the device level. Affective computing offers a new degree of cognitive intelligence into edge processing IoT devices by inferring human emotion, stress levels for intelligent human assistance. This work explores the design space and runtime optimization opportunity for affective computing at the system-on-chip (SoC) level. A design optimization methodology for the neural network classifier and runtime power management schemes are proposed to achieve high energy efficiency on embedded low power devices. A test chip based on a 65nm CMOS process was used to demonstrate the proposed methodology on emotion and stress classification for affective computing. An average power saving of 45% is achieved with a peak power savings of 60% from the proposed emotion-driven adaptive power management scheme.

Keywords

Affective Computing; Internet-of-Things; Embedded Device; Stress and Mood Classification; Neural Network Accelerator.

1 INTRODUCTION

Ultra-low power Internet-of-Things (IoT) or wearable embedded devices have become one of the fastest-growing industry segments. According to a survey from Cisco, the number of devices has sustained an exponential growth rate and will soon reach 50 billion connected devices [1]. While low power and low cost have been the critical masks for IoT devices, the recent rapid development of artificial intelligence (AI) brings a new level of challenges to such devices, i.e. how to create more intelligence into the resource-limited edge devices. From the hardware point of view, the incorporation of machine learning operation into the embedded system has provided a strong boost to the capability of edge processing IoT or wearable devices. Many commercial IoT products have already had built-in artificial intelligence into devices such as smart home products from Nest Labs or similar emerging companies [2]. At chip level, a rapid development happens on integrating machine learning accelerators into the IC chips to facilitate the support of AI from hardware devices [3]. However, the resource limitation and extremely low power budget have become the bottleneck of the development of intelligent edge devices. As a result, dedicated design methodology and intelligent power management schemes are strongly needed for such devices to support resource and power consuming operation of machine learning techniques, which is the focus of this paper.

A class of commonly used edge devices is the wearable medical device where human's daily activities and health conditions are actively monitored for interactive human assistance [4]. While numerous human daily activities and health indicators such as heart rate, blood pressure, calorie intake etc. have already been tracked from such wearable devices, there is currently a lack of dedicated hardware support for the detection of human emotion. As a matter of fact, human emotion or mood provides a rich amount of information on human cognition. Hence comprehension of human emotion serves as a gateway towards the new generation of artificial intelligence [5]. To better understand and manage human emotion, the so-called affective computing techniques are gaining more and more attention to the development of advanced AI-empowered devices.

Affective computing refers to the study and development of intelligent systems and devices that can recognize, interpret and classify human affects which include both short-term stress, emotion and long-term personality, depression, etc. [6]. Due to the nature of the highly interdisciplinary study, affective computing requires close collaboration among computer science, engineering, psychology and as a result, a low power IoT technique that integrates heterogeneous physiological sensors and machine learning capability becomes an enabling technology to support the development of affective computing [7].

Many prior studies were demonstrated using off-shelf physiological sensors to track human emotion. MIT media labs demonstrated that commercial wearable devices can be used to detect a variety of human affects. For instance, the driver's stress was detected using a combination of electromyogram (EMG), electrocardiogram (ECG), respiration and skin conductance (SC) sensors. The detection of stress helped manage driver's driving behaviors to reduce the chances of accidents [8]. People's happiness was also inferred from a combination of SC, accelerometer data as well as person's location history achieving an accuracy of around 70% [9]. Furthermore, combining with cell phone activities, people's emotions for the next day can be predicted with high fidelity [10]. To better model people's stress, a new stress model "cStress" was proposed based on ECG, respiration and accelerometer data showing 90% accuracy compared with self-reported stress [11]. Besides, based on physiological ECG detection, a just-in-time stress intervention scheme was also proposed by Microsoft to mitigate workplace stress through evaluation of employee's stress and mental load [12]. Recently, as virtual reality (VR) becomes a new venue for home entertainment and online business, affective computing offers an alternative path to track users' experience. For instance, an interactive gaming system was proposed to adaptively change gaming scenes and levels of difficulty based on the gamer's emotion. In such an application, a fast emotion tracking at a scale of seconds was delivered from EEG and ECG measurements to support the highly dynamic activity change in VR gaming [13]. The application of emotion and stress classification spans beyond daily activity tracking. A personality detection system was developed based on monitoring the physiological signal response from the users when watching video clips [14]. A web browsing user experience was monitored based on user's pupil dilation to web content [15].

Despite the above growing popularity in using wearable devices for affective computing, there is a lack of discussion from the hardware perspective on how to design and manage the devices for efficient computing. In fact, almost all the works above relies on machine learning techniques, e.g. artificial neural network (ANN), support vector machine (SVM), decision tree (DT) to perform emotion classification. Due to the lack of machine learning support on existing IoT devices, almost all the above work was based on online sensing but offline classification from PC or smartphone rendering major limitation of affective computing for wearable devices. Due to the lack of existing study on how to develop energy efficient ASIC chip for affective computing, this paper, to the best of our knowledge, for the first time, performs a systematic study on design space and power, accuracy, performance tradeoff in designing machine learning empowered edge devices for affective computing. The contributions of this work are summarized as below: (1) from hardware perspective, the design tradeoff between hardware cost, and accuracy is studied with optimization method proposed for implementing machine learning algorithms, e.g. neural network on a chip; (2) a thorough analysis on power consumption of IoT device for affective computing is provided based on real design and usage scenarios showing tradeoff between the power and accuracy; (3) A runtime adaptive power management scheme is proposed to achieve higher power efficiency; (4) A 65nm CMOS test chip was used to demonstrate the proposed adaptive scheme with more than 2X power saving.

2 AFFECTIVE COMPUTING MODEL AND DATABASE

Fig. 1(a) shows the typical system level configuration for affective computing based on physiological signal processing. Various physiological signals such as ECG, EEG, EMG, SC are sensed by low noise amplifiers (LNA) to deliver a large analog signal for later stages. Mixed-signal circuits such as analog to digital conversion (ADC) and feature extraction circuits are used on the sampled physiological signals to reduce the dimensionality of the incoming data. A classifier such as a neural network is used to create final classification results for people's emotions.

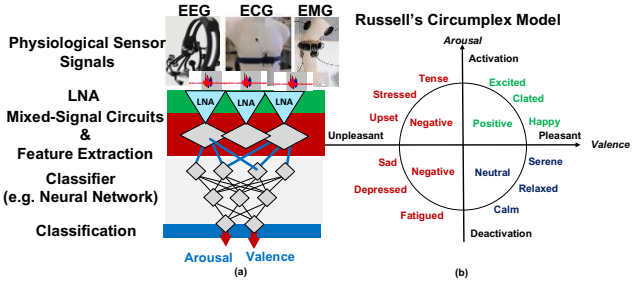


Figure 1: (a) Configuration and signal flow for affective computing; (b) Russell's Circumplex Model [18].

Despite a large variety of applications from affective computing, this work focuses on emotion and stress classification. For stress classification, this work uses the database released from MIT media lab on driver's stress detection from the real-life measurement of drivers' physiological signals [16]. The database provides classification labels of rest, highway and city representing drivers' stress conditions. Although the MIT database only provides a coarse measurement of stress, the driver's stress detection serves as an important application space for affective computing. The requirement for both high accuracy and fast response in driving conditions leads to a strong demand for an efficient wearable device with built-in machine learning capability for fast classification.

To further explore challenges in emotion classification with finer granularity, we also study the DREAMER database which uses off-the-shelf ECG and EEG sensors to detect human's emotion [17]. DREAMER database provides labels on the widely used Russell's Circumplex model for emotion classification [18]. Fig. 1(b) show the Russell's circumplex model where the two-dimensional space of valence and arousal are used to construct people's emotion, such as happiness, upset, and calmness based on the mood angle formed by valence and arousal. Classification accuracy is reported based on the values of valence and arousal. A third variable, dominance is also provided in the database but is not commonly used. In this work, we focus on valence and arousal for emotion inference and use the average accuracy of the two values for accuracy evaluation.

This work targets applications with stress classification for drivers and emotion detection from gaming or virtual reality system. In both cases, fast responses are needed. It has been reported that phasic change of human skin conductance which represents mood swing reaches peak values in 1~5 seconds [15]. Hence, we constrain our classification jobs to be completed within 5 seconds. In different applications, such as emotion detection for online movie recommendation or depression detection, such a requirement can be significantly relaxed.

3 HARDWARE-AWARE CLASSIFIER OPTIMIZATIONS

Although many machine learning algorithms have been explored in previous affective computing studies [17], there is a lack of study considering the dedicated ASIC implementation. In this work, we explore the design space as a tradeoff between power, area, and accuracy. While existing work uses a variety of machine learning schemes such as ANN, SVM, DT, etc. as classifiers, we focus on ANN/DNN in this work due to its high accuracy, scalability, and popularity in the current study.

3.1 Design of Neural Network Classifier

In this work, we used a pipelined multi-layer neural network accelerator as our baseline design [19,20]. As will be shown in our implementation of the SoC chip, the neural network classifier contributes 65% of total area and becomes the largest component of the chip. As a result, it is important to consider the silicon cost when designing the classifier as IoT devices are extremely sensitive to the silicon cost. The optimization of neural network architecture is dictated by the tradeoff between accuracy, silicon area and power consumption. In this study, we vary design parameters such as numbers of neurons in each layer, number of layers to achieve the target accuracy while minimizing power and area cost. 8-bit precision is used in this work. Training is performed offline from PC and weights are downloaded on the SoC chip for classification. As will be shown later, the leakage power consumption of the neural network and SRAM dominates the total digital power consumption due to the long intervals between inference tasks. As a result, the power optimization for the digital classifier is closely related to the optimization of silicon area including neural network size and SRAM spaces. In this work, we assume no power gating is implemented and no non-volatile memory is available to off-load neural network weights.

Fig 2(a) shows the change of accuracy and memory space for the neural network as the number of layers increases in the DREAMER database. The three-layer fully connected neural network is observed to provide the almost-best accuracy. Further increasing layers does not improve the accuracy anymore while incurring more than 20% more memory overhead for each additional layer. The accuracy further degrades beyond four layers due to the difficulty of back-propagation training. As a result, the

optimal number of layers appears to be 3 or 4 layers. Fig 2(b) shows the effect of neuron numbers on accuracy and memory space. As the number of neurons increases, the prediction accuracy increases while the rate of increase starts to saturate after 60 neurons. The amount of memory space needed also increases proportionally with the number of neurons used. As a result, 40~60 neurons appear to be the optimal solution for the application whereas the memory space increases by 34% from 40 to 60 neurons with an accuracy change of 2%. The above study is repeated on the driver stress database showing similar observation and hence is omitted in this work.

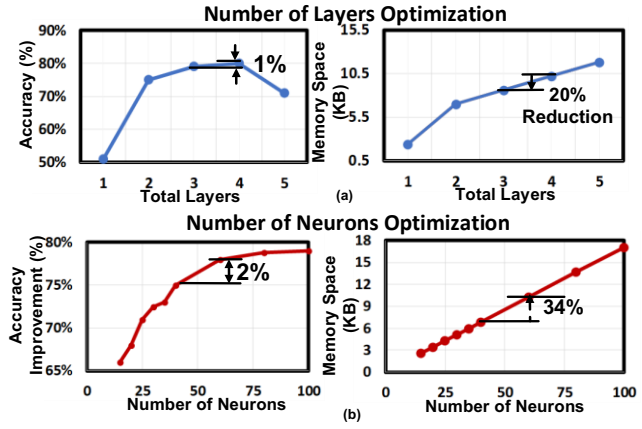


Figure 2: Accuracy and memory space (a) versus the number of layers; (b) versus the number of neurons.

3.2 Optimization of Feature Extraction

Feature extraction also consumes a significant amount of total powers and hence needs to be optimized. In this work, we extract the commonly used time-domain features, e.g. mean, variance, histogram, zero-crossing, slope absolute sign change, to pass to neural network due to the low computation cost of the time-domain features compared with features such as Fourier Transform (FFT), Discrete Wavelet Transform (DWT) which are highly expensive to implement on an edge device.

Fig 3(a) shows the power consumption for generating each feature from each incoming signal assuming a continuous run. For example, the histogram would require the most power to be generated while the zero crossing requires 6X less power. However, feature ranking by power does not consider the significance of each feature for contribution to the final classification. Hence, a more sophisticated ranking methodology needs to be developed to evaluate the importance of the features.

Algorithm 1 Variance-Power Score

Procedure *VP_Score* (*label_list*, *channel_list*, *feature_list*, *feat_power*, *data*)

```
1.  foreach feature ∈ feature_list do
2.    foreach channel ∈ channel_list do
3.      data_s←get_feature(data, channel, feature, sensor)
4.      foreach i ∈ label_list do
5.        foreach j ∈ label_list && j > i do
6.          dist1←extract_distribution(data_s, i)
7.          dist2←extract_distribution(data_s, j)
8.          score_feat(i,j)←ttest (dist1, dist2)
9.        end for
10.     end for
11.     channel_VP_score(channel)←mean(score_feat) *ft_pwr
12.   end for
13. end for
14. return VP_feat_scores  //return the scores
```

To assist in identifying the importance of each feature, we propose a ranking scheme based on a "variance power score", which is the product of "variance" and power for each feature. The "variance" comes from the Kolmogorov-Smirnov statistical test for comparing two different samples. This statistic represents how distinguishable the feature is at various classification labels. The more variance the feature has, the easier for the classifier to perform based on the feature. Algorithm 1 shows the optimization strategy of building a variance-power score for each feature.

In the above algorithm, variance across various groups is compared. Fig. 3(c) shows the variance-power scores across various features. To achieve the highest efficiency, features with lower feature power scores, e.g. variance or slope absolute sign change can be removed to reduce power consumption. Fig 3(b) shows the accuracy tradeoff for removing the least important features across different feature-channel pairs. The accuracy is compared with random removal. As shown in the figure, with 25% removal of features, a 3% accuracy loss and a 25% power saving is observed. Compared with random removal, 3% more accuracy is gained using the proposed ranking methodology. If 13% feature is removed based on the ranking, 20% power saving can be achieved with only 1% accuracy degradation, which represents an optimal power/accuracy tradeoff.

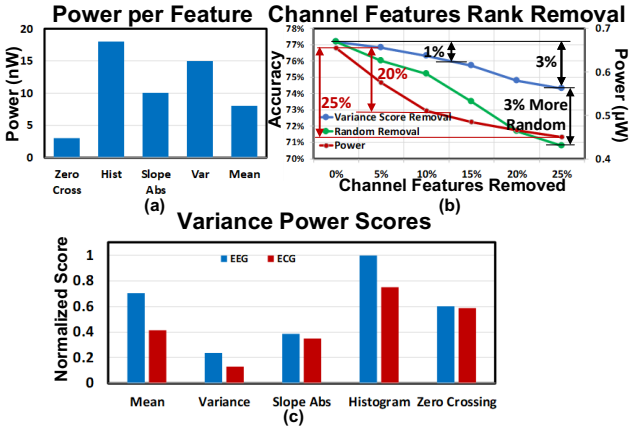


Figure 3: (a) Power comparison among features. (b) Effect of feature reduction based on proposed variance-power score. (c) Variance-power score of each feature.

4 RUNTIME POWER MANAGEMENT

4.1 Operation Modes

Different from high performance computing, physiological signals are slowly varying signals measured at a scale of seconds. It requires high gain low-noise amplifier to amplify micro-volt signal into hundreds of millivolt range to satisfy the input requirement of mixed-signal circuits such as ADC. In the DREAMER database, there are 14 channel EEG signals and 2 channel ECG signals. In the Driver database, there are 1 ECG signal, 1 EMG signal, 2 skin conductance signal, 1 respiration signal as input. The total power consumption was highly related to the working duration of LNA & mixed-signal circuits (MSC). As a result, a new power management paradigm is observed in this work. Fig. 4(a) and Fig. 4(b) shows two different working modes. When chip is working on the continuous mode, LNA & MSC must continuously run for sampling the incoming data. The neural network will be clock gated after a classification is finished. As a result, LNA and MSC dominate total power consumption. To reduce their power consumption, we study a duty-cycle operation mode where the LNA and MSC are only turned on for a fractional period of time. Effectively, the total numbers of raw signal samples are being reduced leading to a drop of accuracy. Our study shows that within a classification interval, e.g. 5 seconds, the classification accuracy strongly depends on the total number of samples in use

but does not strongly depend on when the sampling happened within the classification interval. Hence, the duty cycle directly impacts the final accuracy. Below are the equations for the calculation of total power based on the duty cycle of the operation.

$$DP_{Total} = DP_{NN} \times DC_{NN} + DP_{SRAM} \times DC_{SRAM} + DP_{LNA} \times DC_{LNA} + DP_{MSC} \times DC_{MSC} \quad (1)$$

$$LP_{Total} = LP_{SRAM} + LP_{LNA} + LP_{MSC} \quad (2)$$

$$Power_{Total} = LP_{Total} + DP_{Total} \quad (3)$$

where DP is dynamic power, LP is leakage power, and DC is the duty cycle within the 5-second window.

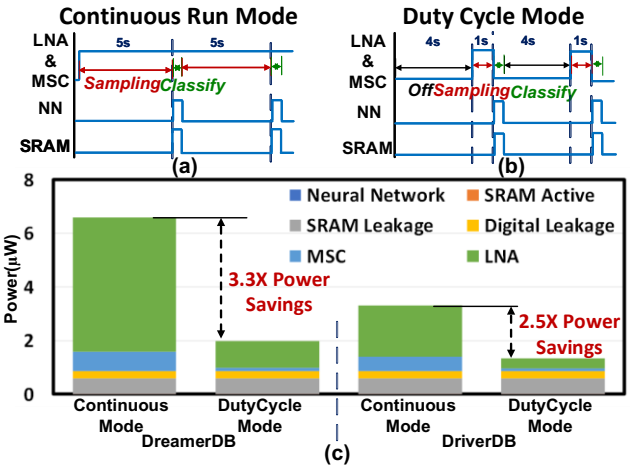


Figure 4: (a) continuous run mode; (b) Duty cycle mode with 20% DC; (c) Power contributions by each circuit on DREAMER and Driver database in two operating modes.

Fig. 4(c) shows the power contribution by different scenarios under two databases on continuous mode and duty cycle mode. The LNA and MSC contribute most power consumption. The leakage power from SRAM and neural network also contribute significantly to the power consumption. It is interesting to observe that in affective computing, the active power from the neural network is less than 5% of the total power. This is because the neural network only requires millisecond to finish classification and remains shut-off most of the time. Compared with continuous run mode, significant power saving is observed due to the duty cycle at an expense of accuracy loss from the shorter sampling time.

4.2 Power and Accuracy Tradeoff

Fig 5 shows the accuracy and power changes as a function of the duty cycle in DREAMER database and Driver database. As the duty cycle decreases, accuracy generally decreases. For instance, at a 50% duty cycle, a 3% accuracy loss is traded off with 30% power saving. Further reduce the duty cycle to 20% can reduce power by 3.3X with 4% loss of accuracy. Based on this observation, we proposed an adaptive power management scheme as will be discussed in section 4.4.

4.3 Proposed Voting Strategy for Accuracy

To help improve prediction accuracy while sustaining lower power of operation, this work proposes a voting strategy. Since the total time for a decision is made at a much slower rate of every 5 seconds, it is possible to vote on the results of multiple times with smaller sampling window. By increasing the number of classifications used to make a final prediction, we are essentially filtering out the number of misclassifications. Equation 4 shows the ideal improvement by using multiple window voting.

$$Majority = \sum_{x=W/2}^{W} \binom{W}{x} p^x (1-p)^{W-x} \quad (4)$$

where p is the prediction accuracy, W is the number of windows to be looked back at. According to the equation, ideally the higher the prediction accuracy is, the greater the improvement is for the voting scheme. As shown in Fig. 6, compared with ideal improvement from (4), the simulation shows a similar trend with lower improvement because samples are temporal correlated and are likely misclassified to the same label. This prevents achieving theoretical voting benefits. As shown in Fig. 6, for both the DREAMER and driver database, the ideal combination would be to use a continuous 1s window with 5 sample voting. Compared to using 5s continuous operation, the accuracy improvement is 2% for the same power consumption. Note that in the voting scenario, the dynamic power of neural network classification increases by 5 times growing from the original 0.1% to 0.4% of the total power, which is still insignificant in total power consumption.

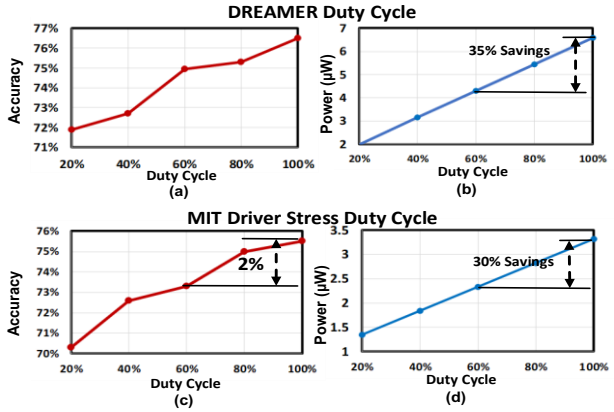


Figure 5: Duty cycle optimization: (a) & (b) DREAMER accuracy vs duty cycle and associated power saving; (c) & (d) MIT Driver Stress accuracy vs duty cycle and associated power saving.

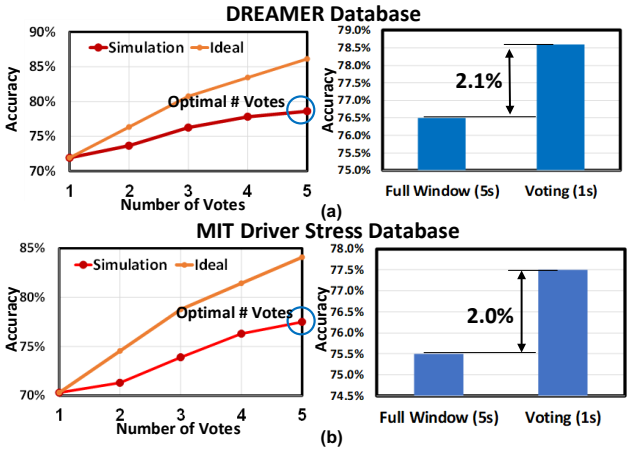


Figure 6: Accuracy with varying number of votes for (a) DREAMER database (b) MIT Driver Stress database.

4.4 Emotion Driven Adaptive Power Management and Sampling

In this section, we propose an adaptive power management and sampling schemes for affective computing based on DREAMER emotion database and MIT driver stress database. The affective computing can be used to manage device power consumption to adapt to human's psychological state during real-life operation, e.g. playing video games or driving. When a sensitive state like fear or high level of stress was detected, the chip can dynamically increase the sampling period to increase the classification accuracy. When a non-sensitive state like calmness was detected, the chip will drop the sampling duty cycle to achieve longer battery life.

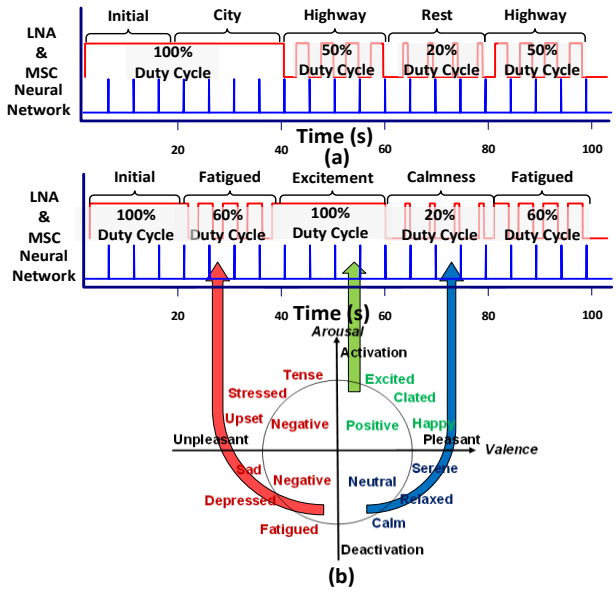


Figure 7: Illustration of proposed emotion driven adaptive power management and sampling scheme for (a) Driving scenario. (b) Gaming scenario.

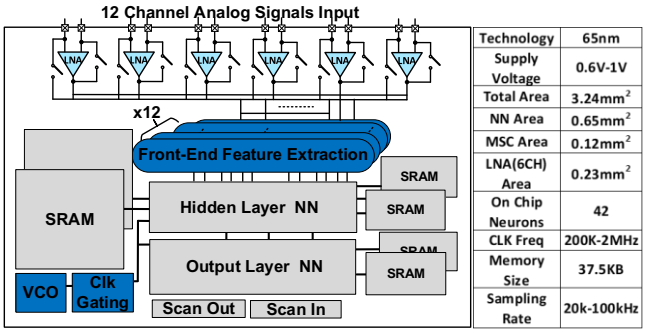
For the DREAMER database, we used a scheme that the video game can dynamically change the difficulty level and environment atmosphere based on emotion state sensed by the chip [14]. A higher level of difficulty and more stressful gaming rhythm will be provided when a low level of emotion like calmness or relax was detected on gamers. At high excitement state, a higher sampling rate is utilized to achieve higher classification accuracy due to the intensity of the actions. At calm state, a lower sampling rate is applied to save power. To perform this task, 6 datasets from DREAMER database with target emotions of excitement, fatigue, and calmness. In the initial state, the chip works on the continuous mode to get the most precise result at the setup phase. When the excitement emotion is detected, the chip works in the continue sampling mode for sensing the gamers emotion condition with high accuracy. When fatigue emotion was detected, the chip will change to 60% duty cycle mode to reduce power and indicate the video game to lower the difficulty and change the environment atmosphere in the game. When the calmness emotion is detected, the chip will be set to a 20% duty cycle to further save the power and indicate the video game to gradually increase the difficulty and intensity of the game to make the video game more challenging.

On driving cases, we set up a scheme that the music of radio channels or A/C temperature can be adaptively changed based on the detected driver's mental stress. We dynamically manage the sampling rate of the device to adapt to the accuracy needed for each driving scenario to achieve better energy efficiency of the IoT device. The driver's stress level is related to different driving environment. The city driving scenario needs much higher attention with a high level of stress measured from the affective metrics. The highway scenario produces less stress level due to the fewer dynamics of the environment. As a result, we propose to dynamically vary the sampling rate to obtain better power and accuracy tradeoff in different scenarios based on the detected driver's stress level. Fig. 7 shows the proposed operation condition. In the initial state, the highest sampling rate, i.e. continuous mode is applied as the driver's stress condition is unknown. As more stress results are detected, based on the different driving scenarios, different duty cycle modes are applied. As shown in Fig. 7, continuous sampling is applied for city condition providing the best accuracy of detection. For highway condition, a 50% duty cycle is applied. At rest condition, a 20% duty cycle is applied. Correspondingly, the accuracy varies from 69% to 75% depending on the stress level of the driver. Many potential applications can be applied with the proposed scheme such as in-vehicle entertainment systems or air-condition controls rending interesting future developments from affective computing.

5 EXPERIMENTAL RESULTS

5.1 Design Overview

To verify the proposed scheme, a 65nm CMOS test chip is fabricated in a low power process as shown in Fig. 8 with design specifications. The chip is designed with up to 12 analog input channels integrating front-end low noise amplifier, mixed-signal data conversion, feature extractions and back-end neural network classifier. All the output features are sent to a neural network classifier with on-chip SRAM cache storing off-line trained weights. Clock-gating is implemented for the neural network and SRAM to dynamically turn on and off the active power.



Technology	65nm
Supply Voltage	0.6V-1V
Total Area	3.24mm^2
NN Area	0.65mm^2
MSC Area	0.12mm^2
LNA(6CH) Area	0.23mm^2
On Chip Neurons	42
CLK Freq	200K-2MHz
Memory Size	37.5KB
Sampling Rate	20k-100kHz

Figure 8: Chip configuration and specification of the test SoC chip.

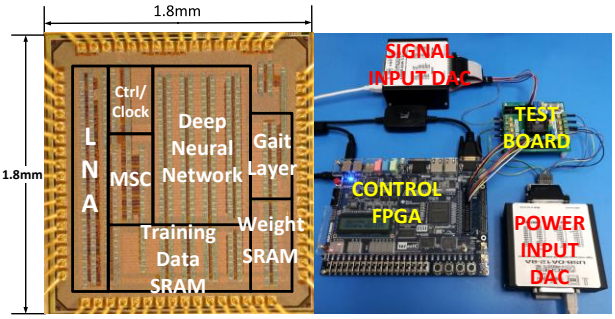


Figure 9: Die micrograph and Test setup.

Fig. 9 shows the die photograph and test setup. The test chip is mounted on a test PCB board. An FPGA was used as an interface for controlling the chip and scanning in and out the data for verification. The selected recorded analog signal channels from MIT Driver Stress database [17] and DREAMER database are replayed with sufficient amplification using the USB-DA12-8A digital to analog converters (DAC) from ACESS. When doing the measurement, due to the limit numbers of LNAs built in the fabricated chip, we selected only five physiological signals for each driver: Electrocardiogram (ECG), Left-shoulder electromyogram (EMG), Left foot skin conductance (SC), Left-hand skin conductance and Chest cavity expansion respiration (RESP) for the Driver database or five EEG signals and one ECG signal from the DREAMER database for the VR gaming case. The chip is operated at the minimum voltage of 0.6V to achieve the lowest power consumption.

5.2 Classification Accuracy and Power Saving

Fig. 10 shows the classification accuracy of the adaptive power management scheme across two database cases. Five seconds windows are used across many samples. In DREAMER cases, the continuously run with excitement emotion achieves 76% accuracy with the highest power consumption while the Calmness emotion

uses a 20% duty cycle with 72% accuracy leading to 60% power reduction. In Driver case, the continuous run achieves 75% accuracy in the city condition, while the duty cycle mode of 50% and 20% are used in the highway and rest respectfully. The accuracy in these cases are 73% and 70%.

Fig. 11 shows the measured sampling & classification waveforms on two databases cases. In the Driver database cases, the chip was set on continues mode on initial state and City state. After Highway state was classified, the chip worked on 50% duty cycle mode, then 20% duty cycle when Rest was sensed. In the Dreamer database case, the initial and excitement state was set in continuous mode. The 20% duty cycle mode was set when Calmness was detected. When fatigue was sensed, the duty cycle was shifted back to 60%.

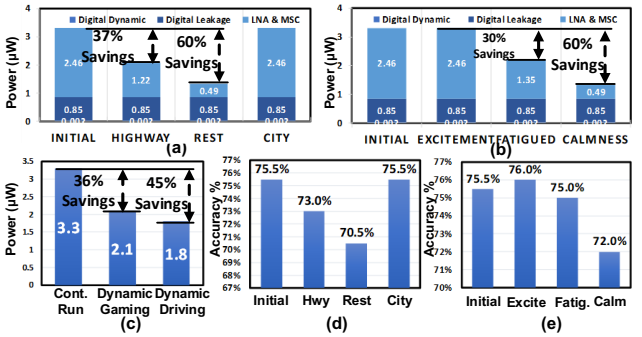


Figure 10: Power and accuracy of measurement. (a) Power in driving scheme; (b) Power in Gaming scheme; (c) Average power saving in proposed schemes; (d) Accuracy in driving scheme; (e) Accuracy in gaming scheme.

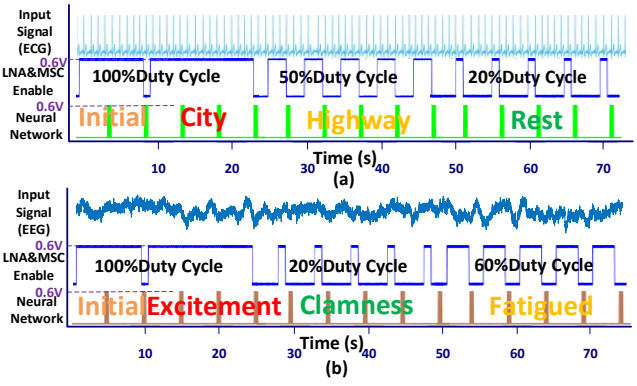


Figure 11: Measured input signal, sampling duty cycle & classification result waveforms in (a) driving scheme, (b) gaming scheme.

Fig. 10 also shows the power measurements break down in various sampling methods. Front end and mixed-signal circuits dominate the total power since it needs a continuous operation. The power consumptions are proportional to the average sampling time in each scenario. The average power using the proposed adaptive power management scheme is reduced by up to 45% compared with the continuous operation mode. A peak power savings of 60% is observed at different stress states of the users. More importantly, compared with conventional dynamic voltage and frequency scaling (DVFS) which does not consider user's affective conditions, the proposed emotion-driven power management provides a new paradigm in device management in the era of artificial intelligence.

6 CONCLUSION

Affective computing provides a new dimension of cognitive intelligence for emerging machine learning empowered edge devices. To study the hardware perspective of affective computing, this paper explores the design space and optimization techniques for designing dedicated ASIC chips. An optimization scheme is proposed to deliver the optimal neural network topography as well as improving the power efficiency of feature extraction. Power management techniques along with voting techniques are proposed to obtain the optimal tradeoff between power consumption and accuracy. An emotion-driven adaptive power management scheme is also proposed to provide runtime optimization for the energy efficiency of affective computing. A 65nm CMOS test chip was used to demonstrate the proposed technique showing 30% to 60% reduction on the power consumption based on the sensed emotion of the users.

ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under grant number CNS-1816870.

REFERENCES

[1] CISCO Online white paper, "The Internet of Things, How the Next Evolution of the Internet is Changing Everything", https://www.cisco.com/c/dam /enus/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf .

[2] Neworkworld, Online Document: "10 Hot AI-powered IoT startups", https://www. networkworld.com/article/3299439/internet-of-things/10-hot-ai-powered-iot-startups.html

[3] Nanalyze, Online Document: "12 AI Hardware Startups Building New AI Chips", https://www.nanalyze.com/2017/05/12-ai-hardware-startups-new-ai-chips/

[4] Mostafa Haghi, et al. "Wearable Devices in Medical Internet of Things: Scientific Research and Commercially Available Devices", *Healthcare Informatics Research*, vol. 23, no. 1, pp. 4-15, 2017.

[5] Forbes, Online Material, "The Next Frontier of Artificial Intelligence: Building Machines That Read Your Emotions", https://www.forbes.com/sites/bernardmarr/2017/12/15/the-next-frontier-of-artificial-intelligence-building-machines-that-read-your-emotions/#2f2035d0647a

[6] Jianhua Tao, Tan Tieniu, "Affective Computing: A Review", *Affective Computing and Intelligent Interaction*, Springer, pp. 981-995, 2005.

[7] Lin Shu, et al. "A Review of Emotion Recognition Using Physiological Signals", *Sensors*, vol. 18, no. 7, 2018.

[8] Jennifer A. Healey, Rosalind W. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors", IEEE Transactions on Intelligent Transportation Systems, vol. 6, no. 2, pp. 156-166, 2005.

[9] Natasha Jaques, et al. , "Predicting Students' Happiness from Physiology, Phone, Mobility, and Behavioral Data", *Affective Computing and Intelligent Interaction*, 2015.

[10] Natasha Jaques, et al. Multi-task Learning for Predicting Health, Stress, and Happiness", *NIPS Workshop on Machine Learning for Healthcare*, 2016.

[11] Karen Hovsepian, et al. "cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment", *Proceeding of International Conference on Ubiquitous Computing*, 2015.

[12] Akane Sano, et al. "Designing Opportune Stress Intervention Delivery Timing using Multi-modal Data," *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017.

[13] Yi Li, et al. "Using Physiological Signal Analysis to Design Affective VR Games", *International Symposium on Signal Processing and Information Technology* , 2015.

[14] Julia Wache, "The Secret Language of Our Body-Affect and Personality Recognition Using Physiological Signals", *International Conference on Multimodal Interaction (ICMI)*, 2014.

[15] Angel Jimenez-Molina, et al. "Using Psychophysiological Sensors to Assess Mental Workload During Web Browsing", *Sensors*, vol. 18, no. 458, 2018.

[16] Jennifer Healey et al , *Driver Stress Data*, http://affect.media.mit.edu, 2002.

[17] Stamos Katsigiannis, et al "DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices", *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98-107, 2018.

[18] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.

[19] Ando, Kota, et al. "BRein Memory: A Single-Chip Binary/Ternary Reconfigurable in-Memory Deep Neural Network Accelerator Achieving 1.4 TOPS at 0.6 W." IEEE Journal of Solid-State Circuits, vol. 53, no. 4, pp. 983–994,2018.

[20] Park, Jeongwoo, et al. "7.6 A 65nm 236.5nJ/Classification Neuromorphic Processor with 7.5% Energy Overhead On-Chip Learning Using Direct Spike-Only Feedback." 2019 IEEE International Solid- State Circuits Conference - (ISSCC), vol. 2019, pp. 140–142, 2019.