frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 13 January 2021
doi: 10.3389/frai.2020.609673

EmotionNet Nano: An Efficient Deep Convolutional Neural Network Design for Real-Time Facial Expression Recognition

James Ren Lee[1]*, *Linda Wang*[1,2] *and Alexander Wong*[1,2]

[1]*Vision and Image Processing Lab, Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada,* [2]*DarwinAI Corp., Waterloo, ON, Canada*

While recent advances in deep learning have led to significant improvements in facial expression classification (FEC), a major challenge that remains a bottleneck for the widespread deployment of such systems is their high architectural and computational complexities. This is especially challenging given the operational requirements of various FEC applications, such as safety, marketing, learning, and assistive living, where real-time requirements on low-cost embedded devices is desired. Motivated by this need for a compact, low latency, yet accurate system capable of performing FEC in real-time on low-cost embedded devices, this study proposes EmotionNet Nano, an efficient deep convolutional neural network created through a human-machine collaborative design strategy, where human experience is combined with machine meticulousness and speed in order to craft a deep neural network design catered toward real-time embedded usage. To the best of the author's knowledge, this is the very first deep neural network architecture for facial expression recognition leveraging machine-driven design exploration in its design process, and exhibits unique architectural characteristics such as high architectural heterogeneity and selective long-range connectivity not seen in previous FEC network architectures. Two different variants of EmotionNet Nano are presented, each with a different trade-off between architectural and computational complexity and accuracy. Experimental results using the CK + facial expression benchmark dataset demonstrate that the proposed EmotionNet Nano networks achieved accuracy comparable to state-of-the-art FEC networks, while requiring significantly fewer parameters. Furthermore, we demonstrate that the proposed EmotionNet Nano networks achieved real-time inference speeds (e.g., >25 FPS and >70 FPS at 15 and 30 W, respectively) and high energy efficiency (e.g., >1.7 images/sec/watt at 15 W) on an ARM embedded processor, thus further illustrating the efficacy of EmotionNet Nano for deployment on embedded devices.

Keywords: face, expression, classification, real-time, neural network

OPEN ACCESS

Edited by:
Fabrizio Riguzzi,
University of Ferrara, Italy

Reviewed by:
Stefano Melacci,
University of Siena, Italy
Marco Lippi,
University of Modena and Reggio Emilia, Italy

***Correspondence:**
James Ren Hou Lee
jrhlee@uwaterloo.ca

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

Received: *23 September 2020*
Accepted: *20 November 2020*
Published: *13 January 2021*

Citation:
Lee JR, Wang L and Wong A (2021) EmotionNet Nano: An Efficient Deep Convolutional Neural Network Design for Real-Time Facial Expression Recognition. Front. Artif. Intell. 3:609673. doi: 10.3389/frai.2020.609673

1. INTRODUCTION

Facial expression classification (FEC) is an area in computer vision that has benefited significantly from the rapid advances in machine learning, which has enabled data collections comprising a diversity of facial expressions captured of different individuals to be leveraged to learn classifiers for differentiating between different facial expression types. In particular, deep learning when applied to FEC has led to significant improvements in accuracy under complex conditions, such as varying lighting, angle, or occlusion.

Even though the performance of deep learning-based FEC systems continue to rise, widespread deployment of such systems is limited, with one of the biggest hurdles being the high architectural and computational complexities of the deep neural networks that drive such systems. This hurdle is particularly limiting for real-time embedded scenarios, where low latency operation is required on the low-cost embedded devices. For example, in the area of assistive technologies for improving quality of life, the majority of individuals using such technologies are unwilling to carry large, bulky, and expensive devices with them during their daily lives, as that would be a big hindrance that limits their ability to leverage the technologies in a seamless manner. As such, the assistive devices must leverage small, low-cost, embedded processors, yet provide low latency to enable real-time feedback to the user. Another example is in-car driver monitoring (Jeong and Ko, 2018), where a FEC system would record the driver and determine their current mental state, and warn them if their awareness level is deteriorating. In cases such as these, the difference of a few milliseconds of processing is paramount for the safety of not only the user, but also other drivers on the road. In applications for fields such as marketing or security, real-time processing is important to provide salespeople or security guards immediate feedback such that an appropriate response can be made as soon as possible. For those relying on software assistance for social purposes, information is required at no delay in order to keep a conversation alive and not cause discomfort for both parties.

A variety of deep neural network architectures have been proposed for FEC, ranging from deep convolutional neural networks (DCNN) to recurrent neural networks (RNN) (Fan et al., 2016) to long-short term memory (LSTM) (Sun et al., 2016) and have been explored, but those introduced in literature have generally required significant architectural complexity and computational power in order to detect and interpret the nuances of human facial expressions. As an alternative to deep learning, strategies leveraging other machine learning strategies such as Support Vector Machines (SVM) (Michel and El Kaliouby, 2003) and hand-crafted features such as Local Binary Patterns (LBP) (Shan et al., 2005; Happy et al., 2012), dense optical flow (Bargal et al., 2016), Histogram of Oriented Gradients (HOG) (Kumar et al., 2016), or Facial Action Coding System (Ekman and Friesen, 1978) have also been explored in literature, but generally have been shown to achieve lower accuracy when compared to deep learning-based approaches, which can better learn the subtle differences that exist between human facial expressions.

To mitigate the aforementioned hurdle and improve widespread adoption of powerful deep learning-driven approaches for FEC in real-world applications, a key direction that is worth exploring is the design of highly efficient deep neural network architectures tailored for the task of real-time embedded facial expression recognition. A number of strategies for designing highly efficient architectures have been explored. One strategy is reducing the depth of the neural network architecture (Khorrami et al., 2015) to reduce computational and architectural complexity; more specifically, neural networks with a depth of just five were leveraged to learn discriminating facial features. Another strategy is reducing the input resolution of the neural network architecture, with Shan et al. (Shan et al., 2005) showing that FEC can be performed even at low image resolutions of 14×19 pixels, which can further reduce the number of operations required for inference by a large margin. Despite the improved architectural or computational efficiencies gained by leveraging such efficient network design strategies, they typically lead to noticeable reductions in facial expression classification accuracy and as such alternative strategies that enable a better balance between accuracy, architectural complexity, and computational complexity are highly desired.

More recently, there has been a focus on human-driven design principles for efficient deep neural network architecture design, ranging from depth-wize separable convolutions (Chollet, 2017) to Inception (Szegedy et al., 2015) macroarchitectures to residual connections (He et al., 2016). Such design principles can substantially improve FEC performance while reducing architectural complexity (Pramerdorfer and Kampel, 2016). However, despite the improvements gained in architectural efficiency, one challenge with human-driven design principles is that it is quite time consuming and challenging for humans to hand-craft efficient neural network architectures that are tailored for specific applications such as FEC that possesses a strong balance between a high performance accuracy, fast inference speed, and low memory footprint, primarily due to the sheer complexity of neural network behaviors under different architectural configurations.

In an attempt to address this challenge, neural architecture search (NAS) strategies have been introduced to automate the model architecture engineering process by finding the maximally performing network design from all possible network designs within a search space. However, given the infinitely large search space within which the optimal network architecture may exist in, significant human effort is often required in designing the search space in a way that reduces it to a feasible size, as well as defining a search strategy that can run within desired operational constraints and requirements in a reasonable amount of time. Therefore, a way to combine both human-driven design principles and machine-driven design exploration is highly desired and can lead to efficient architecture designs catered specifically to FEC.

Motivated by the desire to design deep neural network architectures catered for real-time embedded facial expression recognition, in this study we explore the efficacy of leveraging a human-machine collaborative design strategy that leverages

human experience and ingenuity with the raw speed and meticulousness of machine driven design exploration, in order to find the optimal balance between accuracy and architectural and computational complexity. The resulting deep neural network architecture, which we call EmotionNet Nano, is specifically tailored for real-time embedded facial expression recognition and created via a two phase design strategy. The first phase focuses on leveraging residual architecture design principles to capture the complex nuances of facial expressions, while the second phase employed machine-driven design exploration to generate the final tailor-made architecture design that achieves high architectural and computational efficiency while maintaining a high performance.

To the best of the author's knowledge, the proposed EmotionNet Nano is the very first deep neural network architecture for facial expression recognition leveraging machine-driven design exploration in its design process. All previously introduced FEC network architectures in research literature have been hand-crafted deep neural network architectures with highly uniform architectural characteristics. As a result, the proposed EmotionNet Nano exhibits unique architectural characteristics such as high architectural heterogeneity and selective long-range connectivity not seen in previous FEC network architectures from research literature. We present two variants of EmotionNet Nano, each with a different trade-off between accuracy and complexity, and evaluate both variants on the CK+ (Lucey et al., 2010) benchmark dataset against state-of-the-art facial expression classification networks.

2. MATERIALS AND METHODS

In this study, we present EmotionNet Nano, a highly efficient deep convolutional neural network architecture design for the task of real-time facial emotion classification for embedded scenarios. EmotionNet Nano was designed using a human-machine collaborative strategy in order to leverage both human experience as well as the meticulousness of machines. The human-machine collaborative design strategy leveraged to create the proposed EmotionNet Nano network architecture design is comprised of two main design stages: 1) principled network design prototyping, and 2) machine-driven design exploration.

2.1. Principled Network Design Prototyping

In the first design stage, an initial network design prototype, φ, was designed using human-driven design principles in order to guide the subsequent machine-driven exploration design stage. In this study, the initial network design prototype of EmotionNet Nano leveraged residual architecture design principles (He et al., 2016), as it was previously demonstrated to achieve strong performance on a variety of recognition tasks. More specifically, the presence of residual connections within a deep neural network architecture have been shown to provide a good solution to both the vanishing gradient and curse of dimensionality problems. Residual connections also enable networks to learn faster and easier, with little additional cost to architectural or computational complexity. Additionally, as the network architecture depth increases, each consecutive layer should perform no worse than its previous layer due to the identity mapping option. As a result, residual network architecture designs have been shown to work well for the problem of FEC (Khorrami et al., 2015; Hasani and Mahoor, 2017; Zhou et al., 2019). In this study, the final aspects of the initial network design prototype, φ, consists of an average pooling operation followed by a fully connected softmax activation layer to produce the final expression classification results. The final macroarchitecture and microarchitecture designs of the individual modules and convolutional layers of the proposed EmotionNet Nano were left to the machine-driven design exploration stage to design in an automatic manner. To ensure a compact and efficient real-time model catered toward embedded devices, this second stage was guided by human-specified design requirements and constraints targeting embedded devices possessing limited computational and memory capabilities.

2.2. Machine Driven Design Exploration

Following the initial human-driven network design prototyping stage, a machine-driven design exploration stage was employed to determine the macroarchitecture and microarchitecture designs at the individual module level to produce the final EmotionNet Nano. In order to determine the optimal network architecture based on a set of human defined constraints, generative synthesis (Wong et al., 2018) was leveraged for the purpose of machine-driven design exploration. Defined in **Eq. 1**, we can formulate generative synthesis as a constrained optimization problem, where the goal is to find a generator $\mathcal{G}$ that, given a set of seeds $\mathcal{S}$, can generate networks $\{\mathcal{N}_s | s \in \mathcal{S}\}$ that maximize a universal performance function $\mathcal{U}$ while also satisfying constraints defined in an indicator function $1_r(\cdot)$,

$$\mathcal{G} = \max_{\mathcal{G}} \mathcal{U}(\mathcal{G}(s)) \text{ subject to } 1_r(\mathcal{G}(s)) = 1, \forall s \in \mathcal{S} \qquad (1)$$

As such, given a human-defined indicator function $1_r(\cdot)$ and an initial network design prototype φ, generative synthesis is guided toward learning generative machines that generate networks within the human-specified constraints. The aforementioned constrained optimization problem in **Eq. 1** is solved via an iterative optimization process, where progressively better generator solutions $\widehat{\mathcal{G}}$, as measured based on $\mathcal{U}$, are found while meeting the constraints defined in the aforementioned indicator function $1_r(\cdot)$. The interesting aspect about leveraging an iterative strategy to solve the unconstrained optimization problem is that different generators are found along the way during the optimization process, with the ability to generate deep neural network architectures with different trade-offs between architectural and computational complexity and accuracy.

An important factor in leveraging generative synthesis for machine-driven design exploration is to define the operational constraints and requirements based on the desired task and scenario in a quantitative manner via the indicator function $1_r(\cdot)$. In this study, in order to learn a compact yet highly

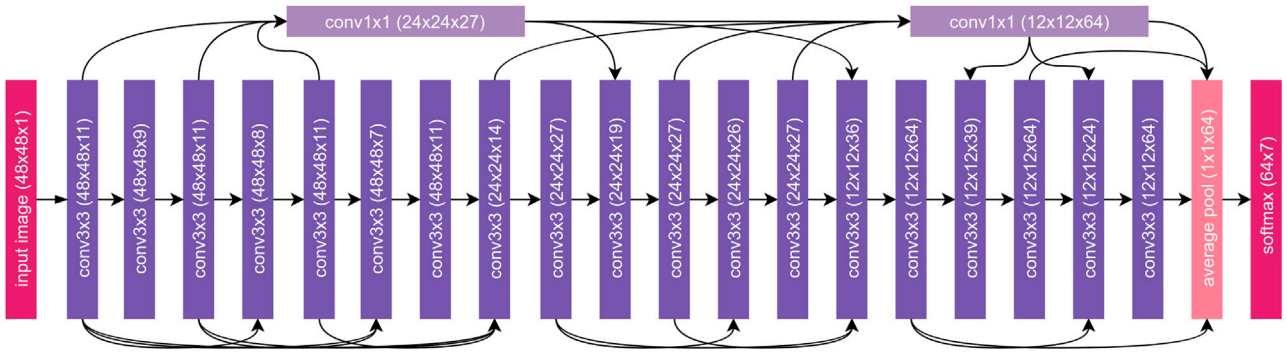

FIGURE 1 | EmotionNet Nano Architecture. The network architecture exhibits high macroarchitecture and microarchitecture heterogeneity, customized toward capturing deep facial features. Furthermore, the network architecture exhibits selective long-range connectivity throughout the network architecture. The number of channels per layer are based on EmotionNet Nano-B.

efficient facial expression classification network architecture, the indicator function $1_r(\cdot)$ was set up such that: 1) accuracy ≥92% on CK+ (Lucey et al., 2010), and 2) network architecture complexity ≤ 1 M parameters. These constraint values were chosen to explore how compact a network architecture for facial expression classification can be while still maintaining sufficient classification accuracy for use in real-time embedded scenarios. As such, we use the accuracy of Feng & Ren (Feng and Ren, 2018) as the reference baseline for determining the accuracy constraint in the indicator function.

The network architecture of the proposed EmotionNet Nano is shown in **Figure 1**. A number of notable characteristics of the proposed EmotionNet Nano network architecture design are worth discussing as they give insights into architectural mechanisms that strike a strong balance between complexity and accuracy.

2.3. Architectural Heterogeneity

A notable characteristic about the architecture that allows the network to achieve high efficiency even with a low number of parameters is the macroarchitecture and microarchitecture heterogeneity. Unlike hand-crafted architecture designs, the macroarchitecture and microarchitecture designs within the EmotionNet Nano network architecture as generated via machine-driven design exploration differ greatly from layer to layer. For instance, there are a mix of convolution layers with varying shapes and different number of channels per layer depending on the needs of the network. As shown in **Figure 1**, there are a greater number of channels needed as the sizes of feature maps decrease.

The benefit of high microarchitecture and macroarchitecture heterogeneity in the EmotionNet Nano network architecture is that it enables different parts of the network architecture to be tailored to achieve a very strong balance between architectural and computational complexity while maintaining model expressiveness in capturing necessary features. The architectural diversity in EmotionNet Nano demonstrates the advantage of leveraging a human-collaborative design strategy as it would be difficult for a human designer, or other design exploration methods to customize a network architecture to the same level of architectural granularity.

2.4. Selective Long-Range Connectivity

Another notable characteristic of the EmotionNet Nano network architecture is that it exhibits selective long range connectivity throughout the network architecture. The use of long range connectivity in a very selective manner enables a strong balance between model expressiveness and ease of training, and computational complexity. Most interesting and notable is the presence of two densely connected 1 × 1 convolution layers that take in outputs from multiple 3 × 3 convolution layers as input, with its output connected farther down at later layers. Such a 1 × 1 convolution layer design provides dimensionality reduction while retaining salient features of the channels through channel mixing, thus further improving architectural and computational efficiency while maintaining strong model expressiveness.

2.5. Dataset

To evaluate the efficacy of the proposed EmotionNet Nano, we examine the network complexity, computational cost and classification accuracy against other facial expression classification networks on the CK+ (Lucey et al., 2010) dataset, which is the most extensively used laboratory-controlled FEC benchmark dataset (Pantic et al., 2005; Li and Deng, 2018).

The Extended Cohn-Kanade (CK+) (Lucey et al., 2010) dataset contains 593 video sequences from a total of 123 different subjects, ranging from 18 to 50 years of age with a variety of genders and heritage. Each video shows a facial shift from the neutral expression to a targeted peak expression, recorded at 30 frames per second (FPS) with a resolution of either 640 × 490 or 640 × 480 pixels. Out of these videos, 327 are labeled with one of seven expression classes, anger, contempt, disgust, fear, happiness, sadness, and surprise. The CK + database is widely regarded as the most extensively used laboratory-controlled FEC database available, and is used in the majority of facial expression classification methods (Li and Deng, 2018; Pantic et al., 2005). **Figure 2** shows that the CK + dataset has good diversity for each

FIGURE 2 | Diversity of expressions in the CK + dataset. Example faces for each expression type in CK+ is shown. Contempt not included as relevant subjects did not give publication consent.

TABLE 1 | Comparison of facial expression classification networks on the CK + dataset. We report 10-fold cross-validation average accuracy on the CK + dataset with seven classes (anger, contempt, disgust, fear, happiness, sadness, and surprise).

Method	Params (M)	Accuracy (%)
Ouellet (2014)	58	94.4
Feng and Ren (2018)	332	92.3
Wang and Gong (2019)	5.4	97.2
Otberdout et al. (2019)	11	98.4
MobileNetV1 (Howard et al., 2017)	3.23	74.6
EfficientNetB0 (Tan and Le, 2020)	4.06	75.9
ResNet50 (He et al., 2016)	23.6	83.4
EmotionNet Nano-A	0.232	97.6
EmotionNet Nano-B	0.136	92.7

expression type, which is important from an evaluation perspective. However, as the CK + dataset does not provide specific training, validation, and test set splits, a mixture of splitting techniques can be observed in literature. For experimental consistency, we adopt the most common dataset creation strategy where the last three frames of each sequence is extracted and labeled with the video label (Li and Deng, 2018). In this study, we performed subject-independent 10-fold cross validation on the resulting 981 facial expression images.

2.6. Implementation Details

EmotionNet Nano was trained for 200 epochs using an initial learning rate of 1e − 3, multiplied by 1e−1, 1e − 2, 1e − 3, and 0.5e − 3 at epochs 81, 121, 161, and 181 respectively. Categorical cross-entropy loss was used with the Adam (Kingma and Ba, 2014) optimizer. Data augmentation was applied to the inputs, including rotation, width and height shifts, zoom, and horizontal flips. Following this initial training, we leveraged a machine-driven exploration stage to fine tune the network specifically for the task of FEC. Training was performed using a GeForce RTX 2080 Ti GPU. The Keras (Chollet, 2015) library was leveraged for this study.

3. RESULTS

3.1. Performance Evaluation

Two variants of EmotionNet Nano were created to examine the different trade-offs between architectural and computational complexity and accuracy. In order to demonstrate the efficacy of the proposed models in a quantitative manner, we compare the performance of both variants against state-of-the-art facial expression classification networks introduced in literature, shown in **Table 1**. It can be observed that both EmotionNet Nano-A and Nano-B networks achieve strong classification accuracy, with EmotionNet Nano-A in particular achieving comparable accuracy with the highest-performing state-of-the-art networks that are more than a magnitude larger. While EmotionNet Nano-B has lower accuracy than the highest-performing networks, it is still able to achieve comparable accuracy as (Feng and Ren, 2018) while being three orders of magnitude smaller with regards to the number of parameters. A more detailed discussion of the performance comparison will be provided in the next section; overall, it can be observed that both EmotionNet Nano variants provide the greatest balance between accuracy and complexity, making it well-suited for embedded scenarios.

The distribution of expressions in CK+ is unequal, which results in an unbalanced dataset both for training and testing. The effects of this are prevalent when classifying the contempt or fear expressions, both of which are underrepresented in CK+ (e.g. there are only 18 examples of contempt, whereas there are 83 examples of surprise). Due to the nature of human facial

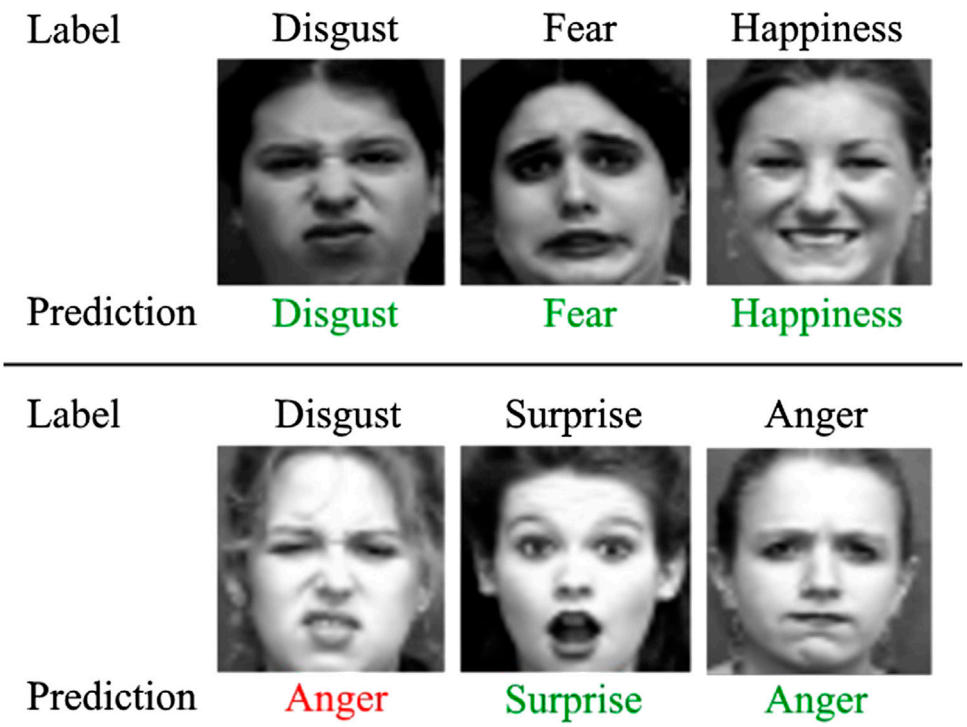

FIGURE 3 | Example expression predictions of faces in the CK + dataset using EmotionNet Nano-A. Five of the faces are classified correctly, indicated in green, with an example of a misclassified expression (disgust), shown in red.

TABLE 2 | EmotionNet Nano Speed and Energy Efficiency. All metrics are computed on an ARM v8.2 64-Bit RISC embedded processor at different power levels.

	15 W		30 W	
Model	**FPS**	$\left[\frac{\text{images/s}}{\text{watt}}\right]$	**FPS**	$\left[\frac{\text{images/s}}{\text{watt}}\right]$
EmotionNet Nano-A	25.8	1.72	70.1	2.34
EmotionNet Nano-B	32.8	2.19	72.9	2.43

expressions, similarities between expressions do exist, but the networks are generally able to learn the high-level distinguishing features that separate one expression from another. However, incorrect classifications can still occur, as shown in **Figure 3**, where a "disgust" expression is falsely predicted to be "anger."

3.2. Speed and Energy Efficiency

We also perform a speed and energy efficiency analysis, shown in **Table 2**, to demonstrate the efficacy of EmotionNet Nano in real-time embedded scenarios. Here, an ARM v8.2 64-Bit RISC embedded processor was used for evaluation. Referring to **Table 2**, both EmotionNet Nano variants are able to perform inference at >25 FPS and >70 FPS on the tested embedded processor at 15 and 30 W respectively, which more than adequately fulfills a real-time system constraint. In terms of energy efficiency, both EmotionNet Nano variants demonstrated high power efficiency, with the Nano-B variant running at 2.43 images/sec/watt on the embedded processor.

4. DISCUSSION

In this study, we explore the human-machine collaborative design of a deep convolutional neural network architecture capable of performing facial expression classification in real-time on embedded devices. It is important to note that other extremely fast deep convolutional neural network architectures exist, such as MicroExpNet (Çuğu et al., 2017), which is capable of processing 1851 FPS on an Intel i7 CPU, is less than 1 MB in size, and is tested on the CK+ 8 class problem (7 facial expression classes plus neutral) on which it achieves 84.8% accuracy. Although a motivating result, a direct comparison cannot be made with EmotionNet Nano as well as other facial expression classification networks evaluated in this study due to the different class sizes.

Compared against state-of-the-art facial expression classification network architectures tested on CK + using the same seven expression classes (see **Figure 1**), both variants of the proposed EmotionNet Nano are at least an order of magnitude smaller yet provide comparable accuracy to state-of-the-art network architectures. For example, EmotionNet Nano-A is $>23\times$ smaller than (Wang and Gong, 2019), yet achieves higher accuracy by 0.4%. Furthermore, while EmotionNet Nano-A achieves an accuracy that is 0.8% lower than the top-performing network architecture (Otberdout et al., 2019), it possesses $>47\times$ fewer parameters. In the case of EmotionNet Nano-B, it achieved higher accuracy (by 0.4%) than (Feng and Ren, 2018) while having three orders of magnitude fewer parameters. Although EmotionNet Nano-B had a lower accuracy than (Ouellet, 2014) by 1.7%, it possesses over 400 times fewer parameters than that model as well.

FIGURE 4 | Assistive technology for Autistic Spectrum Disorder. Example of how EmotionNet Nano can be leveraged to assist individuals with Autistic Spectrum Disorder to better infer emotional states from facial expressions during social interactions in the form of augmented reality.

When compared against well-known deep neural network architectures such as ResNet50 (He et al., 2016), as well as recent state-of-the-art efficient deep neural network architectures such as MobileNetV1 (Howard et al., 2017) and EfficientNetB0 (Tan and Le, 2020), both EmotionNet Nano variants achieve accuracies that is significantly higher while having significantly lower computational complexities (e.g., EmotionNet Nano-A achieves >16% greater accuracy than EfficientNetB0 while having >17 × fewer parameters), showcasing the power of the human-machine collaborative design strategy that is able to create architectures tailor made for the task of expression classification.

Looking at the experimental results around inference speed and energy efficiency on an embedded processor at different power levels (see **Table 2**), it can be observed that both variants of EmotionNet Nano achieved real-time performance and high energy efficiencies. For example, in the case of EmotionNet Nano-A, it was able to exceed 25 FPS and 70 FPS at 15 W and 30 W, respectively, with energy efficiencies exceeding 1.7 images/s/watt and 2.34 images/s/watt at 15 W and 30 W, respectively. This demonstrates that the proposed EmotionNet Nano is well-suited for high-performance facial expression classification in real-time embedded scenarios. An interesting observation that is worth noting is the fact that while the inference speed improvements of EmotionNet Nano-B over EmotionNet Nano-A exceeds 27% at 15 W, there is only a speed improvement of 4% at 30 W. As such, it can be seen that EmotionNet Nano-B is more suitable at low-power scenarios but at high-power scenarios the use of EmotionNet Nano-A is more appropriate given the significantly higher accuracy achieved.

4.1. Implications and Concerns

The existence of an efficient facial expression classification network of running in real-time on embedded devices can have an enormous impact in many fields, including safety, marketing, and assistive technologies. In terms of safety, driver monitoring or improved surveillance systems are both areas that benefit from higher computational efficiency, as it lowers the latency between event notifications as well as reduces the probability that a signal will be missed. With a real-time facial expression classification system in the marketing domain, companies will gain access to enhanced real-time feedback when demonstrating or promoting a product, either in front of live audiences or even in a storefront. The largest impact however, is likely in the assistive technology sector, due to the increased accessibility that this efficiency provides. The majority of individuals do not have access to powerful computing devices, nor are they likely to be willing to carry a large and expensive system with them as it would be considered an inconvenience to daily living.

As shown in this study, EmotionNet Nano can achieve accurate real-time performance on embedded devices at a low power budget, granting the user access to a facial expression classification system on their smartphone or similar edge device with embedded processors without rapid depletion of their battery. This can be extremely beneficial toward tasks such as depression detection, empathetic tutoring, or ambient interfaces, and can also help individuals who suffer from Autistic Spectrum Disorder better infer emotional states from facial expressions during social interaction in the form of augmented reality (see **Figure 4** for a visual illustration of how EmotionNet Nano can be used to aid in conveying emotional state via an augmented reality overlay).

Although EmotionNet Nano has many positive implications, there exist concerns that must be considered before deployment. The first concern is privacy, as individuals may dislike being on camera, even if no data storage is taking place. Privacy concerns, especially ones centered around filming without consent, are likely to arise if these systems start to be used in public areas. The combination of facial expression classification together with facial recognition could result in unwanted targeted advertising, even though this could be seen as a positive outcome for some. Additionally, wrong classifications could result in unintended implications. When assisting a user in an ambient interface or expression interpretation task, a misclassified expression could result in a negative experience with major consequences. For example, predicting "sad" or "angry" expressions as "happy" could influence the user to behave in the wrong manner. These concerns and issues are all worth further exploration

and investigation to ensure that such systems are used in a responsible manner.

5. CONCLUSION

In this study, we introduced EmotionNet Nano, a highly efficient deep convolutional neural network design tailored for facial expression classification in real-time embedded scenarios by leveraging a human-machine collaborative design strategy. By leveraging a combination of human-driven design principles and machine-driven design exploration, the EmotionNet Nano architecture design possesses several interesting characteristics (e.g., architecture heterogeneity and selective long-range connectivity) that makes it tailored for real-time embedded usage. Two variants of the proposed EmotionNet Nano network architecture design were presented, both of which achieve a strong balance between architecture complexity and accuracy while illustrating performance trade-offs at that scale. Using the CK + dataset, we show that the proposed EmotionNet Nano can achieve comparable accuracy to state-of-the-art facial expression classification networks (at 97.6%) while possessing a significantly more efficient architecture design (possessing just 232 K parameters). Furthermore, we demonstrated that EmotionNet Nano can achieve real-time inference speed on an embedded processor at different power levels, thus further illustrating its suitability for real-time embedded scenarios.

Future work involves incorporating temporal information into the proposed EmotionNet Nano design when classifying video sequences. Facial expressions are highly dynamic and transient in nature (Hasani and Mahoor, 2017), meaning that information about the previous expression is valuable when predicting the current expression. Therefore, the retention of temporal information can lead to increased performance, at the expense of computational complexity. Investigating this trade-off between computational complexity and improved performance when leveraging temporal information in combination with machine-driven designs would be worthwhile.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.pitt.edu/~emotion/ck-spread.htm. The CK+ dataset was used for this study.

AUTHOR CONTRIBUTIONS

JL and AW contributed to conception and design of the study. JL assembled and organized the database. JL and LW ran the experiments and performed statistical analysis. JL wrote the first draft of the manuscript. LW and AW contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was partially supported by Microsoft.

ACKNOWLEDGMENTS

The authors would like to thank NSERC, Canada Research Chairs program, and Microsoft.

REFERENCES

Bargal, S. A., Barsoum, E., Ferrer, C. C., and Zhang, C. (2016). "Emotion recognition in the wild from videos using images," in Proceedings of the 18th ACM international conference on multimodal interaction. (ACM), Tokyo, Japan, 433–436.

Çuğu, İ., Şener, E., and Akbaş, E. (2017). Microexpnet: an extremely small and fast model for expression recognition from frontal face images. Preprint repository name [Preprint]. Available at: arXiv:1711.07011 (Accessed November 19 2017).

Chollet, F. [Dataset] (2015). *Keras*. https://keras.io (Accessed March 27 2015).

Chollet, F. (2017). "Xception: deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, October 2016. Available at: arXiv:1610.02357v3 (Accessed April 04 2017) [abstract]. 1251–1258.

Ekman, P., and Friesen, W. V. (1978). *Facial action coding system: a technique for the measurement of facial manual*. Palo Alto, CA: Consulting Psychologists Press.

Fan, Y., Lu, X., Li, D., and Liu, Y. (2016). "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in Proceedings of the 18th ACM international conference on multimodal interaction. (ACM), Tokyo, Japan, 445–450.

Feng, D., and Ren, F. (2018). "Dynamic facial expression recognition based on two-stream-cnn with lbp-top," in 2018 5th IEEE international conference on cloud computing and intelligence systems (CCIS), Nanjing, China, November 23–27, 2018 (IEEE), 355–359.

Happy, S., George, A., and Routray, A. (2012). "A real time facial expression classification system using local binary patterns," in 2012 4th International conference on intelligent human computer interaction (IHCI), Kharagpur, India, December 27–29, 2012 (IEEE), 1–5.

Hasani, B., and Mahoor, M. H. (2017). "Facial expression recognition using enhanced deep 3d convolutional neural networks;" in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, Honolulu, HI, November 21–26, 2017 (IEEE), 30–40.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, June 27–30, 2016 (IEEE), 770–778.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: efficient convolutional neural networks for mobile vision applications. Preprint repository name [Preprint]. Available at: arXiv:1704.04861 (Accessed April 17 2017).

Jeong, M., and Ko, B. C. (2018). Driver's facial expression recognition in real-time for safe driving. *Sensors* 18, 4270. doi:10.3390/s18124270

Khorrami, P., Paine, T., and Huang, T. (2015). "Do deep neural networks learn facial action units when doing expression recognition?" in Proceedings of the IEEE international conference on computer vision workshops, Santiago, Chile, December 7–13, 2015 (IEEE), 19–27.

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. Preprint repository name [Preprint]. Available at: arXiv:1412.6980 (Accessed December 23 2017).

Kumar, P., Happy, S., and Routray, A. (2016). "A real-time robust facial expression recognition system using hog features," in 2016 International conference on

computing, analytics and security trends (CAST), Pune, India, December 19–21, 2016 (IEEE), 289–293.

Li, S., and Deng, W. (2018). *Deep facial expression recognition: a survey*. Preprint repository name [Preprint]. Available at: arXiv:1804.08348 (Accessed April 23 2018).

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). "The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression," in 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, San Francisco, CA, June 13–18, 2010 (IEEE), 94–101.

Michel, P., and El Kaliouby, R. (2003). "Real time facial expression recognition in video using support vector machines," in Proceedings of the 5th international conference on multimodal interfaces. (ACM), Vancouver, BC, 258–264.

Otberdout, N., Kacem, A., Daoudi, M., Ballihi, L., and Berretti, S. (2019). Automatic analysis of facial expressions based on deep covariance trajectories. *IEEE Trans. Neural Networks Learn. Syst.* 31, 3892–3905. doi:10.1109/TNNLS.2019.2947244

Ouellet, S. (2014). Real-time emotion recognition for gaming using deep convolutional network features. Preprint repository name [Preprint]. Available at: arXiv:1408.3750 (Accessed Auguest 16 2014).

Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). "Web-based database for facial expression analysis," in 2005 IEEE international conference on multimedia and Expo, San Francisco, CA, July 6, 2005 (IEEE), 5.

Pramerdorfer, C., and Kampel, M. (2016). *Facial expression recognition using convolutional neural networks: state of the art*. Preprint repository name [Preprint]. Available at: arXiv:1612.02903 (Accessed December 09 2016).

Shan, C., Gong, S., and McOwan, P. W. (2005). "Recognizing facial expressions at low resolution," in IEEE conference on advanced video and signal based surveillance, 2005, Como, Italy, September 15–16, 2005 (IEEE), 330–335.

Sun, B., Wei, Q., Li, L., Xu, Q., He, J., and Yu, L. (2016). "LSTM for dynamic emotion and group emotion recognition in the wild," in Proceedings of the 18th ACM international conference on multimodal interaction. (ACM), Tokyo, Japan, 451–457.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MU, June 7–12, 2015 (IEEE), 1–9.

Tan, M., and Le, Q. V. [Dataset] (2020). *Efficientnet: rethinking model scaling for convolutional neural networks*. Preprint repository name [Preprint]. Available at: arXiv:1905.11946v5 (Accessed May 28 2020).

Wang, G., and Gong, J. (2019). "Facial expression recognition based on improved lenet-5 cnn," in 2019 Chinese control and decision conference (CCDC), Nanchang, China, June 3–5, 2019 (IEEE), 5655–5660.

Wong, A., Shafiee, M. J., Chwyl, B., and Li, F. (2018). *Ferminets: learning generative machines to generate efficient neural networks via generative synthesis*. Preprint repository name [Preprint]. Available at: arXiv:1809.05989 (Accessed September 17 2018).

Zhou, Y., Ren, F., Nishide, S., and Kang, X. (2019). "Facial sentiment classification based on resnet-18 model," in 2019 International conference on electronic engineering and informatics (EEI), Nanchang, China, June 8–10, 2019 (IEEE), 463–466.

Conflict of Interest: LW and AW are affiliated with DarwinAI Corp.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.