Virtual Assistant Attention Detection

Ethan Callanan[1], Jae Makitalo[2], Ella Duffy[3], Kavin Zhu[4]

QMIND – Queen's AI Hub
Queen's University, Kingston, Ontario K7L 3N6, Canada. Queen's

1 e-mail: e.callanan@queensu.ca
2 e-mail: jae.makitalo@queensu.ca
3 e-mail:17emld@queensu.ca
4 e-mail:18kz44@queensu.ca

Abstract: *The popularity of virtual assistants has been rising at an exponential rate, but they all use the same unnatural method of keyword activation. The goal of this project was to develop a novel system to provide a more natural interface for interacting with virtual assistant devices. To achieve this, we developed an attention detection system using a multitask cascaded convolutional neural network for face detection and a convolutional neural network for attention classification. The face detector performs with a true positive rate of 95.04%, and the attention classifier performs with 97.2% testing accuracy. The attention detection pipeline was implemented in a web application simulating a virtual assistant. We plan on improving the generalizability of the attention classifier by training it on a larger and more diverse dataset, and we plan on implementing the model in a dedicated device.*

1. INTRODUCTION

1.1 Motivation

The use of virtual assistants (VA) has seen a meteoric rise in past years. In the last two years alone, the number of VAs in use worldwide has risen from 3.25 billion to 4.2 billion. By 2024 that number is projected to overtake the world population with approximately 8.4 billion devices [1]. The text-to-speech recognition segment of the VA market alone was valued at USD 2.2 billion in 2019, and the market is expected to grow at a rate of 34.4% over 2020 to 2027 [2]. Despite the technology's incredible popularity, the way users interact with the devices has not seen any development. In social interactions, humans naturally focus their attention on the speaker; however, none of the major devices implement vision based interaction and instead opt for unnatural keyword activation

1.2 Related Works

Developers in the VA field have begun incorporating computer vision in their products for applications unrelated to activation. Google has implemented gesture controls and uses facial recognition for personalized display in their Nest Hub Max, and Amazon utilizes face detection to orient the Echo Show towards the user. Previous applications of attention detection have largely focused on driver monitoring. Although they are not designed for VAs, they operate on the same principles. The most notable implementation is Comma AI's driver monitoring system, which utilizes eye tracking and image classification to determine whether or not the driver is paying attention to the road. Researchers at the Massachusetts Institute of Technology (MIT) built a gaze estimation model for driver monitoring which avoids the use of eye tracking [3]. Instead, they opted to perform face detection with a Histogram of Oriented Gradients and linear support vector machine to detect faces, extract the facial landmarks with a cascade of regressors from a facial landmark mark-up, and classify the gaze direction in one of six regions with a random forest classifier.

1.3 Problem Definition

In recent years, the market and use of VAs has grown rapidly, but the way in which we interact with these assistants has been largely overlooked. We set out to design a novel method of interaction, using computer

vision, to provide a user experience that more closely resembles that of a normal conversation. As mentioned in *Section 1.2*, using computer vision for tasks related to attention detection has been explored before. CommAI's driver monitoring system heavily relies on eye tracking, however, eye detection is unreliable in this application due to the many angles and lighting conditions a user may interact with the device at. MIT proposes a more suitable implementation, however their system uses a six stage pipeline and is a six class classifier. For the purposes of VA activation, binary classification will suffice and is both simpler and less computationally expensive.

2. METHODOLOGY

2.1 Dataset Generation

Training and evaluation is carried out on a dataset of 10 subjects. For each subject, there are 200 real images and 324 synthetic images, providing 5240 total images [4]. The images are varied in illumination, background, and pose (by up to 30 degrees in either direction). This dataset was supplemented with an additional 200 images of a sitting subject with similar variety. Preprocessing of the images involved converting to grayscale and resizing to 224 pixels along the smallest edge (maintaining the aspect ratio). Each image was labelled as either "attentive" or "inattentive" based on whether or not the subject was looking towards the camera. A 20% test split was used to evaluate the models.

2.2 Solution

The solution consists of a two step pipeline: face detection and attention classification. If the system passes the first step (face detection) the attention classifier is activated and makes the binary decision as to whether or not the user is focusing their attention on the device.

2.3 Face Detection

A multitask cascaded convolutional neural network [5, 6] (MTCNN) was used to identify if a face is present in an image frame. The network consists of three stages in the form of independent convolutional neural networks (CNN).

The first stage, the proposal network, uses a fully convolutional network[1] (FCN). This network finds windows in the image that could potentially contain a face as bounding box regression vectors. The network performs some refinement to combine overlapping regions, and outputs the remaining candidate windows. Next, the refine network performs calibration with bounding box regression and uses non-maximum suppression to further combine overlapping windows. It then outputs whether each candidate contains a face or not, along with a bounding box and vector for facial landmark localization (eyes, nose, and mouth).Finally, the output network operates in a similar fashion to the refine network, but describes the face in more detail. This final stage outputs the binary face classification, along with the bounding box and five absolute landmark locations: the two eyes, nose, and mouth corners.

2.4 Attention Classification

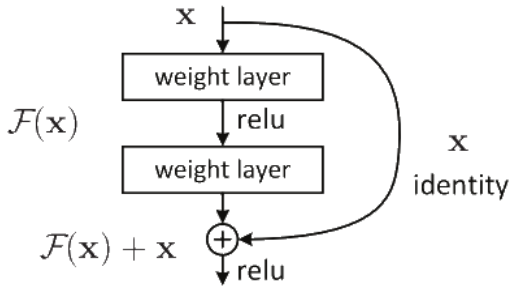

Figure 1: A residual block in the ResNet architecture. Layers can skip subsequent layers in the network through an identity shortcut connection.

A CNN was used to make a binary classification on the attentive state of a face. The classifier uses the ResNet [7] architecture with 50 convolutional layers. The network achieves far better results with less training than its shallower counterparts, and manages to avoid the problem of vanishing gradients[2] by introducing identity shortcut connections. These connections allow a layer to skip the subsequent layers and map its output directly to a layer further in the network as shown in figure 1. The first layer is a 7×7 kernel, the second layer is a 3×3 max pool, and each subsequent layer is a 3×3 kernel, all using rectified linear unit activation. Dropout was applied for regularization and to prevent co-adaptation of neurons.

[1] A CNN without a dense layer.

[2] Repeated multiplication during backpropagation causes the gradient to shrink. If a network is sufficiently deep, this will cause massive degradation in performance.

Training was performed with the Adam optimizer [8] using negative log likelihood loss for 10 epochs.

2.5 Virtual Assistant Integration

The model was implemented in a Streamlit web-application made to simulate a VA device. The model analyzes every other frame to make its classification on the user's attentive state. When the user is attentive for 10 consecutive frames (five consecutive positive classifications from the model), the app waits for the user to begin speaking and listens until they complete their sentence. The recording is then sent to a custom Dialogflow agent through the dialogflow API and both the audio and text responses are displayed to the user.

3. RESULTS AND DISCUSSION

The face detector performs at a true positive rate of 95.04% and the attention classifier achieved an accuracy of 97.2% on the test set.

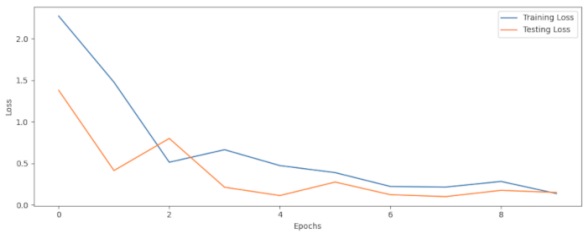

Figure 2: Learning curve of the attention classifier.

The accuracy of the models closely reflects its practical performance in conditions similar to the training data. Interacting with the assistant was a nearly seamless user experience[3], and false positives were handled by the assistant activation logic in the application. False negatives from the classifier occasionally delay the activation of the assistant, but these occurrences are infrequent enough not to diminish the overall user experience.

Despite the attention classifier's accuracy in controlled conditions, when presented with poor lighting or unfamiliar angles the performance suffered. This is likely largely due to the consistent set of lighting and angles in the training images. Additionally, all the images were taken at similar distances from the camera. As such, in significantly unfamiliar conditions the model will get stuck on one of the two classifications.

4. CONCLUSIONS AND FUTURE WORK

The two components of the attention detection pipeline were successfully built. Both models performed well in a testing environment and in controlled live environments. In unfamiliar contexts the attention classifier did not perform as well. Reflection on the training data suggests this was due to insufficient variety in the image attributes. The models were integrated with a proof of concept VA application and provided a positive user experience.

We aim to improve the generalizability of the attention classifier by training on a more varied dataset. Images of subjects taken from different angles, elevations, and distances will help the model handle the many edge cases that arise from live classification. Performance in poor light conditions may also be improved by adding more images in low light and with different light sources. Finally, we plan on implementing the model and activation logic in a dedicated VA device.

REFERENCES

[1] Statista Research Department, "Number of voice assistants in use worldwide 2019-2024", Statista, 22 Jan 2021.

[2] Grand View Research, "Intelligent Virtual Assistant Market Size, Share & Trends Report", Grand View Research, Apr 2020.

[3] L. Fridman, P. Langhans, J. Lee, B. Reimer, "Driver Gaze Region Estimation Without Using Eye Movement", MIT, 1 Mar 2016.

[4] "Face Recognition Database", MIT Center for Biological and Computational Learning.

[5] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks", IEEE Signal Processing Letters, 11 Apr 2016.

[6] https://github.com/davidsandberg/facenet

[7] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.

[3] Demo interaction is shown at https://www.youtube.com/watch?v=0-YFEVMPsV8