Studying and Recreating Misactivations on a Google Home Mini

Nicole Gerzon

This project was motivated by the work done by a similar study [1] studying the misactivations of a range of home speakers, as well as a secondary study [2] that did a full evaluation on the mini's capabilities. The goal was to test if the findings of the first study could be recreated after nearly a year of mini updates and software bug fixes, and if the findings would be different when taking advantage of Google's option to store a particular voice and train the mini to respond better to that voice.

The original paper used audio from a collection of TV shows, which was effective in maintaining a consistent data set with readily available captions, but not realistic for the average use case of a smart speaker. I opted to record and use around 30 hours of audio from around my apartment so it would include a more balanced sampling of what the mini would be expected to hear on a day-to-day basis. The audio was trimmed for any places of dead noise and was exclusively representative of actual conversations. The audio was collected throughout the course of a week and if I was to do it again I would increase the data set by recording for a longer period of time. This negative became apparent as the experiment progressed and it was clear that misactivations don't necessarily trigger in reproducible ways.

For collecting the data to prove the existence of an activation of the google mini, I used Wireshark to collect data on network traffic, as well as Google's "My Activity" cloud server which tracks the usage of different google devices. The google server sorts mini activations by device and includes details such as the wake word it believes was used, the query to the mini, and sometimes the location of the query. In cases where a question is not asked, the activation is titled "Unknown Voice Command" or "Assistant Used". For the purposes of this study, I also turned on the option for the server to store the recording of the activations along with the other details.

I ran into similar issues as the study I was following when tracking and recognizing misactivations. Aside from minor issues with reading the network traffic, Google's cloud servers are not entirely indicative of all activations read by the mini. Many of the activations to a wake word (that was not followed by a question) were not noted on Google's cloud interface, despite an evident rise in traffic and having the mini's lights turn on. This leads me to believe that Google doesn't report all possible activations on its server, maintaining a certain threshold of measuring how legitimate/important the activity was before reporting it. Therefore, for the purposes of this paper, I relied mostly on data from the network traffic and used the cloud activations as a supplement for information on what kind of data the mini seems to cache, as well as what is stored as a legitimate activations, versus a mistake.

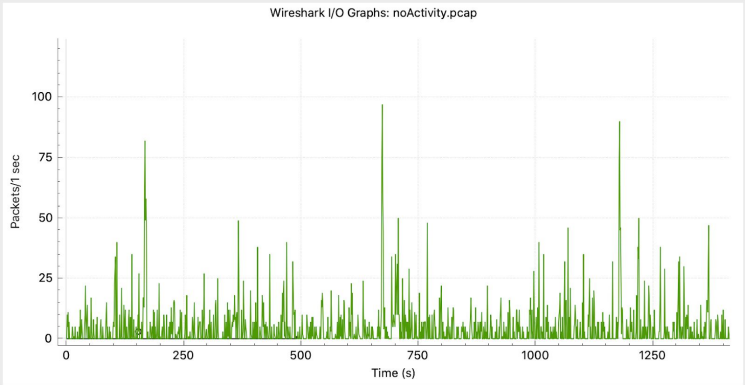


Figure 1: Background Traffic Sample

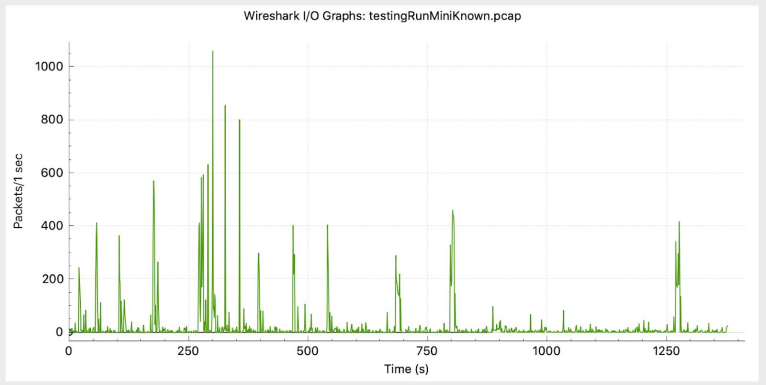


Figure 2: Purposeful Activations

All of the data was collected using Wireshark, having my laptop and the google mini connected to a wifi network on a router that is connected to Northeastern's ethernet. Since the Google Mini automatically communicates with Chrome, while the experiments were run, Chrome was completely closed and the laptop was put into monitor mode to further cut off any disruptions to the readings. The figures track communication with a filter that tracks the MAC address of the mini.

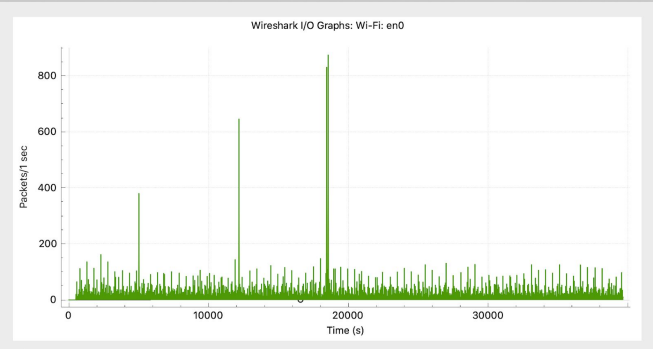
When collecting information on misactivation patterns, the first step was to define a threshold for an activation of the google home mini. This was done by allowing the mini to run undisturbed for a couple hours and come up with a range of typical background traffic. Figure 1 below shows a sample of this background traffic. This continuous activity sends up to 150 packets of data, either to the router as a keep-alive for the connection, or a continuous connection maintenance with the google servers that it queries in the case of an activation. For the remainder of the experiments, any traffic reading below a threshold of 100 was ignored entirely as background traffic.

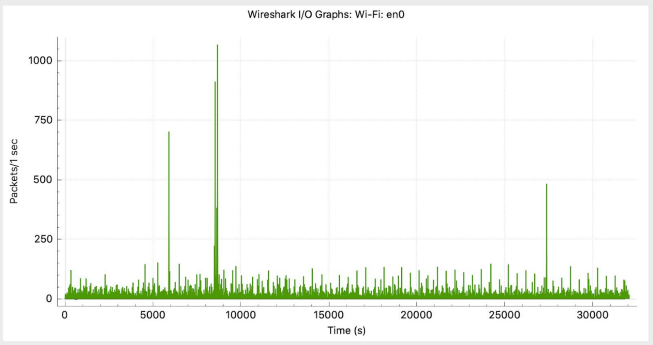
Figure 2 provides clarity on how activation ranges were measured. The Google Mini responds differently depending on the type of query it is faced with. In Figure 2, I attempted to get a rage for the different types of questions, those that require an internet lookup, those that should be cached within the mini and require little to no external help, and simple activations by wake word. This was done by asking a series of questions in succession. The first 4 required offline uses of the home, such as setting an alarm,the next four queried heavy internet use, such as playing music, and the last four were just wake words.

From this experiment and similar recreations, I was able to gather a basic threshold for categorizing any activation I would be seeing. Any communication above 250 packets I considered a wake word activation and anything over ~400 packets was tracked and stored by Google's server, therefore also being considered by Google as a valid activation worth reporting. In general, most of the misactivations I was able to identify as being even partially repeatable were 300 packets and above.

The next steps for the experiment was playing the audio and checking if any of

the activations were large enough to surpass the thresholds defined earlier.





Figures 3&4: Overnight Tests

Both Figures 3 and 4 above represent eight hour snippets of audio play throughs where an activation is easily visible. In the case of both graphs, the largest spike in network traffic can be attributed to what I believe is a daily caching event that was present in all the overnight network traces. In both of these figures, the main point of interest is the activation that sends just under 500 packets of data. These two activations were responses to a mistaken trigger word, either “Ok - cool!” or “Hey guys!!”, which not only triggered an increase in network traffic but was also stored in Google’s cloud server. These two activations were continuously repeatable. Figure 5 is a closeup of one of these repeatable activations.

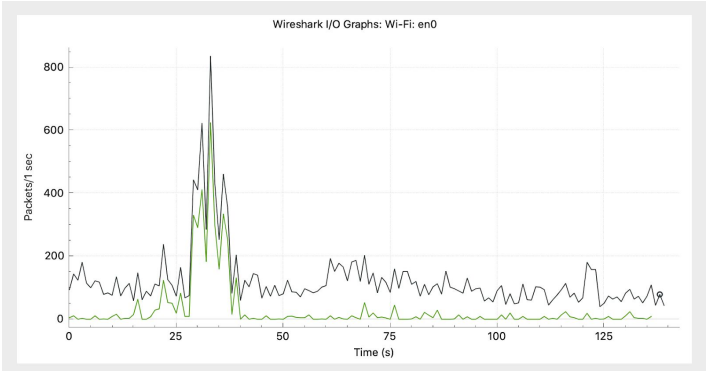


Figure 5: Repeatable Misfire

Other misactivations, while triggering the network traffic minimum to be considered valid, were not able to be reliably repeatable and were not stored on Google’s servers. Furthermore, aside from misheard wake words, the mini had several instances where an activation coincided with a time in the audio where there were multiple conflicting speakers. These misactivations were not recreatable, and seemed almost random.

Originally, I expected to see a decrease in misactivations in most cases, since I expected the mini to focus on the voice it was trained on, however, I couldn’t find a direct correlation between the misactivations and the voice that was speaking. Out of the 30 hours of audio, there were two continuously reproducible misactivations, one tied to the trained speaker. The other misactivations showed up randomly throughout the different experiments which is overall consistent with the original study, which found a low rate of reproducibility in misactivations. Furthermore, since the other misactivations had no correlation with the speaker, overall this shows that activating voice recognition

does not completely solve the issue of misactivations on the Google Mini.

Due to issues with reading and capturing the network traffic at the beginning of the experimentation process, unfortunately I don't have enough data to make generalizations on other trends that I suspect are influencing misactivations such as the presence of an accent or second language, the number of speakers, etc... For the rest of the paper, there will instead be a discussion on other interesting observations I was able to make about the mini, its activities, interactions with your other existing technology, and open questions about Google's information collection and storage that I'm left with at the end of this project.

One of the first things I noticed about Google's servers is that they can be configured to store recordings of activations, and those recordings start before the wake word is said. This means that even if the recordings are being constantly collected in 10 second intervals and then replaced with new audio, the microphone is still actively collecting and storing the data that comes before a wake word, even for a short amount of time. In the case of a misactivation, despite Google servers not leaving a visible trace of a recording, I think it would be interesting to see if the data is still being stored from that misactivation and how the mini decides how much of this stored data to keep and send.

Another interesting finding is that the Google mini is in constant communication with Google chrome if it is open on the same network. When doing a general packet capture, my laptop and the mini were continuously communicating back and forth. This feature is advertised as a way to integrate google apps and other devices together, there is even a Local Home SDK that Google offers to execute chrome javascript apps and activate speakers and other smart devices connected to the same network. The same communication happens between the Assistant and your phone if it opens the google app or Chrome, a connection is immediately established and maintained.

Another thing I found especially interesting when working with the mini is its ability to cache and call back information. I first noticed this quality when running test commands for finding activation thresholds, I couldn't ask the same question twice within a day and get the same result, because the second time didn't warrant an internet search and just registered as a wake word activation. This was more or less expected when it came to repeat questions, but certain queries that were tied to location data, such as the weather, or information tied to my google account were also cached.

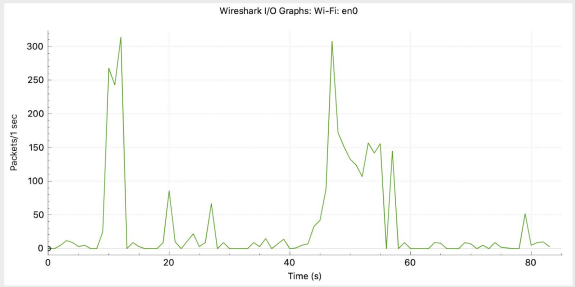


Figure 6: Caught Caching

For example, for the first peak in Figure 6 the question is: "What's the weather like today?" while the second is a simple "Hey Google". Both queries send the same amount of data despite the fact that

one requires a response and data and the other does not. This behavior can be explained by the nightly data pull observed in both Figures 3 and 4, but does not explain how the weather data remains current throughout the day unless the mini is constantly refreshing this storage of information for the most common/expected questions.

Finally, this was something I discovered in the last two days, the Mini isn't well equipped when it comes to understanding foreign languages - but it still tries to respond to foreign queries. When I try to query the mini in Russian, I still get a response, just not to the question I asked. For example, a query such as: "What's the weather like today", will trigger a schedule explanation. I thought this was fascinating because the mini tries its hardest to analyze the sounds but cannot seem to understand that they aren't in English. I'm not sure if this behavior is tied to the specific frequencies of a particular foreign language. Trying in spanish caused a lower amount of misunderstandings, because the mini was able to realize it didn't understand the language but it was still triggered in some instances and gave random responses.

Overall, I found this work to be incredibly interesting and eye-opening, since I wasn't aware of a lot of these habits, despite having a mini of my own for many years. These tests showed that the issue with misfires still readily exists when owning at least the google home assistant and that there are many more unknowns in how Google collects and stores our data.

Sources:

[1] Dubois, Daniel J., et al. *When Speakers Are All Ears: Characterizing Misactivations of IoT Smart Speakers*. moniotrlab.ccis.neu.edu/wp-content/uploads/2020/06/dubois-pets20.pdf.

[2] Park, Minjin, and Joshua I. James. *Preliminary Study of a Google Home Mini*. arxiv.org/pdf/2001.04574.pdf.

[3] Forrester, *Smart Home Devices Forecast*, 2017 To 2022 (US). Accessed on 02/28/2020, https://www.forrester.com/report/Forrester+Data+Smart+Home+Devices+Forecast+2017+To+2022+US/-/E-RES140374.

[4] Artem Russakovskii, *Google is permanently nerfing all Home Minis because mine spied on everything I said 24/7*. Accessed on 02/28/2020, https://www.androidpolice.com/2017/10/10/google-nerfing-home-minis-mine-spiedeverything-said-247/.

[5] VRT NWS, *Google employees are eavesdropping, even in your living room*. Accessed on 02/28/2020, https://www.vrt.be/vrtnws/en/2019/07/10/google-employeesare-eavesdropping-even-in-flemish-living-rooms/.

[6] Google, Google Assistant. Accessed on 02/28/2020, https: //assistant.google.com.

[7]