CHA: A Caching Framework for Home-based Voice Assistant Systems

Lanyu Xu\textsuperscript{1}, Arun Iyengar\textsuperscript{2}, Weisong Shi\textsuperscript{1}

\textsuperscript{1}Wayne State University
\textsuperscript{2}IBM T.J. Watson Research Center
Introduction: Smart Speaker

U.S. Adult Smart Speaker Installed Base January 2020

- **Total US Adult Population**: 255 MILLION
- **Jan 2020**: 87.7 MILLION (32.0% One-Year Growth)
- **Jan 2019**: 66.4 MILLION
- **Jan 2018**: 47.3 MILLION

Status-quo Approach

Cloud-based

Command understanding

action: active
object: light
location: kitchen

API service External APIs Database

Fulfillment

“Turn on the light in the kitchen”

1. audio command

2. audio command

3. detection response

3. detection response

4. response

4. response

5. response

5. response

6. take action

6. take action

[Motivation 1] Command happens in home, fulfills in home.
Limitations

• FAQ collected from Google and Amazon product forums

**Google Home mini** slow to respond to commands
Community forum - Google Nest 1 Recommended Answer
4/24/19 - Looks like the problem was caused by the mini being on 5GHz. Connected it to the 2.4GHz frequency and is back to normal operation.

**Hub is responding slowly, Google Support has been "helping" for a**
Community forum - Google Nest 1 Reply
8/17/19 - Please expect a longer than normal response time as a result of recent current events. We appreciate your patience and understanding as we work to provide ...

**Does ANYONE Here Have Continued Conversations on the GH Hub**
Community forum - Google Nest 11 Replies
2/15/19 - Please expect a longer than normal response time as a result of recent current events. We appreciate ... I think they are probably a slow roll out but let me check.

**Very slow response in light control lately, and problems with**
Community forum - Google Nest 1 Reply
9/25/20 - Please expect a longer than normal response time as a result of recent current events. We appreciate your patience and understanding as we work to provide ...

---

**My Alexa (2nd generation) response time has slowed significantly. Any ideas on how to resolve this issue?**
Echo - DJ5111 - October 7, 2020 at 1:52 PM

**Echo show slow response time**
Echo Show - MrRox - January 27, 2020 at 10:38 PM

**Echo slow, delayed responses, mishearing and deaf not hearing well**
Echo - Egnek - October 24, 2019 at 8:27 PM

**echo stops playing radio, and gives slow response**
Echo - adebyrne - October 4, 2020 at 10:08 AM

**My 2nd generation dot is slow to respond**
Echo Dot - MarionNH - May 18, 2020 at 8:52 PM

---

User Behavior

• Google home usage survey\textsuperscript{[1]}
  • 65,499 utterances, 88 diverse homes, over 110 days.
  • Limited command length: 1 – 10 words, median 4 words.
  • Highly spatial-temporality related:
    • \textasciitilde 3 domains/household.
    • Active usage 7AM – 11PM, peaks 5-6PM.
  • Semantic duplicated: frequently change commands for same information.

\begin{center}
\end{center}
CHA: An overview

1. audio command
2. audio command
3. cache lookup
   - action: active object: light location: kitchen
4. response
5. cache hit, take action
6. detection response
7. response
8. take action

Cloud-based

“Turn on the light in the kitchen”
Contributions

• Identifying two drawbacks of the cloud-based voice assistant system.

• Developing an edge-based caching framework to improve user experience.

• Exploring system efficiency strategies for resource-constraint devices in home environment.
# Experiment Setup

## Cloud

- Dialogflow

## Edge

- Raspberry Pi 4B (b)
- Intel FRD
- Jetson Xavier

## Smart Speaker

- Raspberry Pi 4B (a)
- Mic Array

## Hardware

<table>
<thead>
<tr>
<th>Hardware</th>
<th>CPU</th>
<th>GPU</th>
<th>Memory (GB)</th>
<th>Cost (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raspberry Pi 4B</td>
<td>ARMv7</td>
<td>N/A</td>
<td>4</td>
<td>55</td>
</tr>
<tr>
<td>Intel Fog Reference Design</td>
<td>Intel Xeon E3-1275</td>
<td>N/A</td>
<td>32</td>
<td>N/A</td>
</tr>
<tr>
<td>Jetson AGX Xavier</td>
<td>ARMv8</td>
<td>512-core Volta</td>
<td>32</td>
<td>699</td>
</tr>
</tbody>
</table>
## Dataset

- **Fluent Speech Commands**
  - Typical smart home commands in English: home automation, task management.
  - 1 – 9 words / spoken command.
  - 31 intents, 3 slot types.
  - 4 – 24 types of expressions / intent. 248 unique utterances.

<table>
<thead>
<tr>
<th>Intent (trigger)</th>
<th>Commands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase volume</td>
<td>Louder please.</td>
</tr>
<tr>
<td></td>
<td>Turn sound up.</td>
</tr>
<tr>
<td></td>
<td>I can’t hear that.</td>
</tr>
<tr>
<td></td>
<td>I need to hear this, increase the volume.</td>
</tr>
<tr>
<td>Active kitchen light</td>
<td>Turn on the kitchen light.</td>
</tr>
<tr>
<td></td>
<td>Switch on the kitchen light.</td>
</tr>
<tr>
<td></td>
<td>Kitchen light on.</td>
</tr>
</tbody>
</table>
Cloud-only or Edge-based?

<table>
<thead>
<tr>
<th></th>
<th>Word error rate (WER)</th>
<th>Sentence accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud-only ASR</td>
<td>10.42%</td>
<td>83.19%</td>
</tr>
<tr>
<td>Edge-based ASR</td>
<td>2.52%</td>
<td>96.12%</td>
</tr>
</tbody>
</table>

**Response time (s)**

<table>
<thead>
<tr>
<th>Audio size (KB)</th>
<th>Edge-based ASR</th>
<th>Cloud-based ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>35</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>41</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>43</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>45</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>46</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>48</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>50</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>52</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>55</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>56</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>59</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>61</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>64</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>68</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>71</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>76</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>96</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

ASR: Automatic speech recognition
Cloud-only or Edge-based?

NLU: Natural language understanding

- Edge-based ASR
- Cloud-based ASR
- Cloud-based ASR-NLU

Audio size (KB)

Response time (s)

Edge brings lower latency, more stable performance comparing to cloud-only processing.
System Design

Response latency
Understanding accuracy
System efficiency

“Turn on the light in the kitchen”
Intent (trigger): active_kitchen_light

Hash table
<key: trigger, value: action>

Trigger: “active kitchen light”
Entity: light.kitchen
Status: (state == off)
Action: state.on

Command Interface layer
Message queue

Command parser layer
Adaptive learning
Understanding

Command caching layer
Caching
Rerouting

Management layer
Storage
Configuration

Smart devices (illumination/surveillance/entertainment/…)

RESTful API
Voice Assistant
Command Understanding

• Goal
  • Audio input $\rightarrow$ (intent, slot)

• Methodology
  • Automatic speech recognition + natural language understanding (ASR + NLU)
    • Conventional method
  • Spoken language understanding (SLU)
    • Extracts words and phoneme features
    • followed by intent detection and slot filling

• CHA
  • ASR: pocketsphinx\textsuperscript{[2]}
  • NLU: BERT\textsuperscript{[3]}

<table>
<thead>
<tr>
<th>Turn</th>
<th>On</th>
<th>The</th>
<th>Light</th>
<th>In</th>
<th>The</th>
<th>kitchen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slot</td>
<td>B-active</td>
<td>I-active</td>
<td>O</td>
<td>B-object</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>

Intent

Active_kitchen_light

---


Command Understanding (cont’d)

- Inherit from BERT
  - Pre-trained distilBERT
  - Jointly detect intent and slot types

Improve for cache miss?  Pruning layers

Latency (ms)

<table>
<thead>
<tr>
<th></th>
<th>Cloud</th>
<th>Raspberry Pi</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
<td>1,600</td>
<td>1,200</td>
</tr>
<tr>
<td>NLU</td>
<td></td>
<td>800</td>
</tr>
</tbody>
</table>

Size reduction: 53%  Acceleration: 5.8X
System Efficiency

• Workload
  • Simulate query in Pareto distribution.
    • Probability distribution $f(\text{trigger}, \alpha) = \frac{\alpha}{\text{trigger}^{\alpha+1}}$. Higher $\alpha$ has higher semantic locality.
    • $\alpha = 0.25, 0.5, 1.0$, and uniform distribution.
    • Cache warmup with 5, 10, 20 commands.

• Insight
  • On Raspberry Pi, CHA provides a fast and stable response with a lightweight understanding module.
CHA on Different Edge Devices

• Response time
  • Reduced by 70%, 94%, 77% than the cloud-only solution for cache hit item.
  • Low overhead for cache missed item.

• Resource utilization
  • Low resource consumption across platforms.
  • System loading takes 13, 2, 24 seconds on three platforms, respectively.

• CHA has generality to be deployed on different hardware equipped devices.
Discussion

• Layer pruning benefits BERT and its variants with subtle performance degradation (when pruned to 1 layer).

<table>
<thead>
<tr>
<th></th>
<th>Layers</th>
<th>Model size (MB)</th>
<th>Param size (million)</th>
<th>Intent accuracy</th>
<th>Slot F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>12 → 1</td>
<td>438 → 126</td>
<td>110 → 30</td>
<td>96% → 92%</td>
<td>96.3%</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>6 → 1</td>
<td>256 → 123</td>
<td>66 → 30</td>
<td>92%</td>
<td>96.3%</td>
</tr>
<tr>
<td>ALBERT</td>
<td>1</td>
<td>46.87</td>
<td>12</td>
<td>96%</td>
<td>96.3%</td>
</tr>
</tbody>
</table>

• End-to-end SLU model compression is challenging due to its dense and informative structure (compare to compressed NLU model).

<table>
<thead>
<tr>
<th></th>
<th>Raspberry Pi</th>
<th>Intel FRD</th>
<th>Jetson Xavier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inference time</td>
<td>737.0 ms (127.2 ms)</td>
<td>41.4 ms</td>
<td>83.0 ms</td>
</tr>
<tr>
<td>Model size</td>
<td>15.9 MB (123.8 MB)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter size</td>
<td>3 million (30 million)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion and Future Work

• Conclusion
  • CHA is proposed to address two drawbacks for cloud-based voice assistant systems.
  • CHA integrates a set of compression strategies to provide affordable and practical solution for home-based voice assistant systems.
  • CHA provides a 70% acceleration in voice command processing on the low-cost, resource-constrained raspberry pi, with low resource consumption.

• Future work
  • Exploring audio caching.
  • Developing model compression strategies.
Thank you!

http://thecarlab.org/

xu.lanyu@wayne.edu