Enabling the Next Generation of Smart Devices with Interactive AI

Roland Memisivic
September 2020
Our mission at TwentyBN: Moving from smart speakers...

**Smart speakers** are confined to the simplest of tasks: “how’s the weather?”

Existing “service robots” are essentially chatbots with bodies.
...to embodied assistants, that can see and interact with you

Yoohoo! Over here!

Come up a little higher!

Check these items out!
Millie, an embodied “Alexa” for public spaces
A fitness coach that won’t let you cheat (iPhone 8 and up)

FitnessAllyApp.com
End-to-end learning is taking over the world

- **2010**
  - Audio
  - Neural network
  - Text

- **2012**
  - Pixels
  - Neural network
  - Objects

- **2014**
  - English
  - Neural network
  - French

...
Humanoid assistants push end-to-end learning to its limits!

<table>
<thead>
<tr>
<th>Audio-visual input</th>
<th>Neural network</th>
<th>Audio-visual input</th>
</tr>
</thead>
</table>
| • Understand the scene  
• Understand objects and actions  
• Understand human  
• Behaviour  | • Understand spoken language  
• Generate spoken language  
• Link visual concepts to words (“grounding”)  
• Reason about past and present events  | • Control the assistant’s face  
• Control the assistant’s body  
• Etc |

- Objects
- Actions
- Attributes
- Speech input
- Speech output
- Body control
- ...
How do we get the data for this!?

- We built a worldwide “movie studio” that cranks out up to 45k training videos a day
- To catch every imaginable edge case, videos are recorded and annotated (crowd acting)
- Videos are automatically recorded, re-coded, reviewed, …
- The data collection is constantly evolving according to the AI systems current needs…

- … resulting in >3.5 million videos across thousands of classes and tasks to date
- The network already learned objects, actions, temporal events, counting, captioning, “behaviours”, …

Researchers identify which labels the network needs to learn next

Crowdworkers record, verify and annotate videos
Bounding boxes?
We let the network figure out where to look!
Skeleton models? We let the network figure out how well you’re doing!

SCORE=0.68
ANGLE=45.08 degrees
SPEED=2.40 rep/s (avg speed=1.97, count estimate=45.05)
Temporal segmentation? We let the network figure out what you’re doing, and when!
End-to-end learning makes data sourcing hard, keeping networks simple

**Apple Devices**
- iPhone 8 and 8 Plus (2017)
- iPhone X (2017)
- iPhone XS, XS Max, and XR (2018)
- iPad Air, Mini and Pro (2019)
- iPhone 11 and 11 Pro (2019)
- iPhone SE (2020)

**Android Devices**
- Samsung Galaxy S8 and Note 8 (2017)
- Samsung Galaxy S9 (2018)
- Samsung Galaxy S20 (2019)
- Google Pixel 2, Xiaomi Mi 6 (2017)
- Sony Xperia XZ2, Xiaomi Mi 8 (2018)
- Sony Xperia 1 II, Xiaomi Mi 10 (2019)

All inference can run with 8.5B MAC/sec -> .017 TOPS
End-to-end learning makes data sourcing hard, keeping networks simple

Apple Devices

- iPhone 8 and 8 Plus (2017)
- iPhone X (2017)
- iPhone XS, XS Max, and XR (2018)
- iPad Air, Mini and Pro (2019)
- iPhone 11 and 11 Pro (2019)
- iPhone SE (2020)

Android Devices

- Samsung Galaxy S8 and Note 8 (2017)
- Samsung Galaxy S9 and Note 9 (2018)
- Samsung Galaxy S20 (2019)
- Google Pixel 2, Xiaomi Mi 6 (2017)
- Sony Xperia XZ2, Xiaomi Mi 8 (2018)
- Sony Xperia 1 II, Xiaomi Mi 10 (2019)

All inference can run with 8.5B MAC/sec -> .017 TOPS
Conclusions

• **Personal companions** and camera-enabled assistants are moving from science fiction to reality.

• They’re a gigantic commercial opportunity because they are **personalized** and **sticky**. The more you use them, the better they get.

• They’re also a gigantic research opportunity because they stretch the limits of AI via **audio-visual dialogue** and **grounding**.

• **On-screen companions** can capture a lot of that value since they allow us to bring to bear **end-to-end** learning for most part of the way.

• For more info and access to our developer platform visit: 20bn.com
Resources

Developer platform
• 20bn.com

Fitness Ally app
• www.fitnessallyapp.com

Embodied AI newsletter
• www.embodiedai.co

The “something something” database for learning common sense features
• arxiv.org/abs/1706.04261

Datasets
• https://20bn.com/products/datasets
Me, me!

Questions?