Tiny models with big appetites: Cultivating the perfect data diet
Plumerai people detection

Production-worthy computer vision on Arm Cortex-A SoCs and even 1-Cortex-M microcontrollers.

Come check these demos live at our booth!
How did we do it?

Highly intertwined, need to cover the entire stack

Check out FPGA demo at our booth!

Crucial for tinyML!

Custom architectures and training strategies

Plumerai Data Pipeline
Plumerai Tiny Models
Plumerai Inference Engine
Plumerai Hardware IP Core

World’s fastest inference engine for Arm Cortex-M: 1.6x faster & 2x less RAM than TFLM¹

¹. https://blog.plumerai.com/2021/10/cortex-m-inference-software/
Why not just use public data?

We need to design a good dataset before we can design a good model!

- Bias: must be worth looking at
- Irrelevant context
- Shortcut: blue background, no people!
- Person-centered!
Plumerai dataset

Not person-centered

Nothing to look at...

Need tooling to make the most of your data!
Plumerai infrastructure
Infrastructure overview

- Model architecture
- Training strategies
Infrastructure overview

- Data collection
- Labeling
  - Sample selection
  - Model architecture
  - Training strategies
- Data unit tests
- Data verification & categorization
- Augmentation
- Hardware
  - Preprocessing
  - Post-processing
  - Sensor devices
- Deployment
- Monitoring

Failure analysis
Data curation
Inference engine
Infrastructure overview

- Data collection
- Labeling
- Sample selection
  - Model architecture
  - Training strategies
- Data unit tests
- Data unit curation
- Hardware
- Deployment
- Post-processing
- Data verification & categorization
- Augmentation
- Sensor devices
- Preprocessing
- Monitoring
- Inference engine
- Failure analysis
Data unit tests

Unit Tests Passed: 73 / 84

- empty_desk
- empty_hallway
- empty_office
- person_desk_face
- person_desk_back
- person_sitting
- person_sitting_on_chair
- person_hallway_within_4m
- person_hallway_4_7m
- person_hallway_beyond_7m
- skin_tone_1
- skin_tone_2
- skin_tone_3

Real-life performance

Cannot capture model behavior in a single metric!

Edge cases
Data curation cycle

Need **good data tooling** to:

1. Identify failure cases
2. Link problem to training data:
   - Visual similarity
   - Interactive classification
   - Training influence
   - ...
3. Address problems
   - Correct labels
   - Oversampling
   - Targeted augmentation
   - Supervised Contrastive Learning
   - Add / remove data
   - ...
4. Re-train model

Test-driven development

This is an iterative process!

Link problem to training data:

Quality > quantity
Demo recap

- Tooling allows us to identify problematic images:
  - Similar appearance
  - Automatic classification
  - Influence on training process
Demo recap

- Tooling allows us to identify problematic images:
  - Similar appearance
  - Automatic classification
  - Influence on training process
Demo recap

- Tooling allows us to identify problematic images:
  - Similar appearance
  - **Automatic classification**
  - Influence on training process
Demo recap

• Tooling allows us to identify problematic images:
  • Similar appearance
  • Automatic classification
  • **Influence on training process**
Demo recap

• Tooling allows us to identify problematic images:
  • Similar appearance
  • Automatic classification
  • Influence on training process

• Act on them in real-time!
• Custom tooling: quickly add / explore new AI-assisted approaches
  • Needed for scalable data iteration
  • 30M images in total
  • Terabytes of data!
Conclusion

Production-worthy tinyML requires:

- **Vertical integration!**
- Rapid iteration
- High-quality data

High-quality model!
Questions
Not part of a healthy diet!
Thank you!
Copyright Notice

This presentation in this publication was presented as a tinyML® Summit 2022. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org