tinyML® Summit

Miniature dreams can come true...

March 28-30, 2022 | San Francisco Bay Area

www.tinyML.org

Accelerated Model Deployment with LEIP Recipes

Haya Sridharan,
Product Manager

Latent AI
March 2022

Latent AI Overview

- We develop software to optimize AI models for low power and low latency.
- Founded Dec 2018 as SRI startup spin-off backed by DARPA technologies
- Team of proven startup-veterans, accomplished in AI. 25 FT staff (MS/PhD)
- Offices in Menlo Park, CA and Princeton, NJ

Awards

https://bit.ly/2EVXsBF

Top 100 AI Startup 2021

https://tcrn.ch/34juQee
https://vimeo.com/460028482

Top 60 Edge Computing Companies

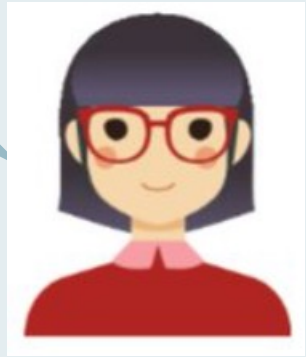
Investors

An Edge ML Story

Let's detect potholes!

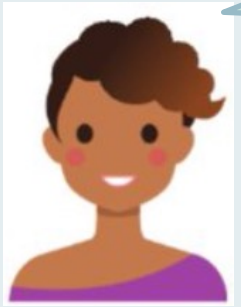
Solutions Steve
(Links with customer)

Ooh! I read a cool new paper, let me train a model :)



Data Scientist
Daisy

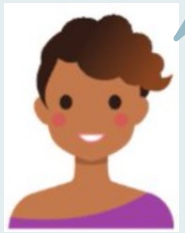
Yikes!!



ML Engineer
Evelyn

Edge ML Challenges

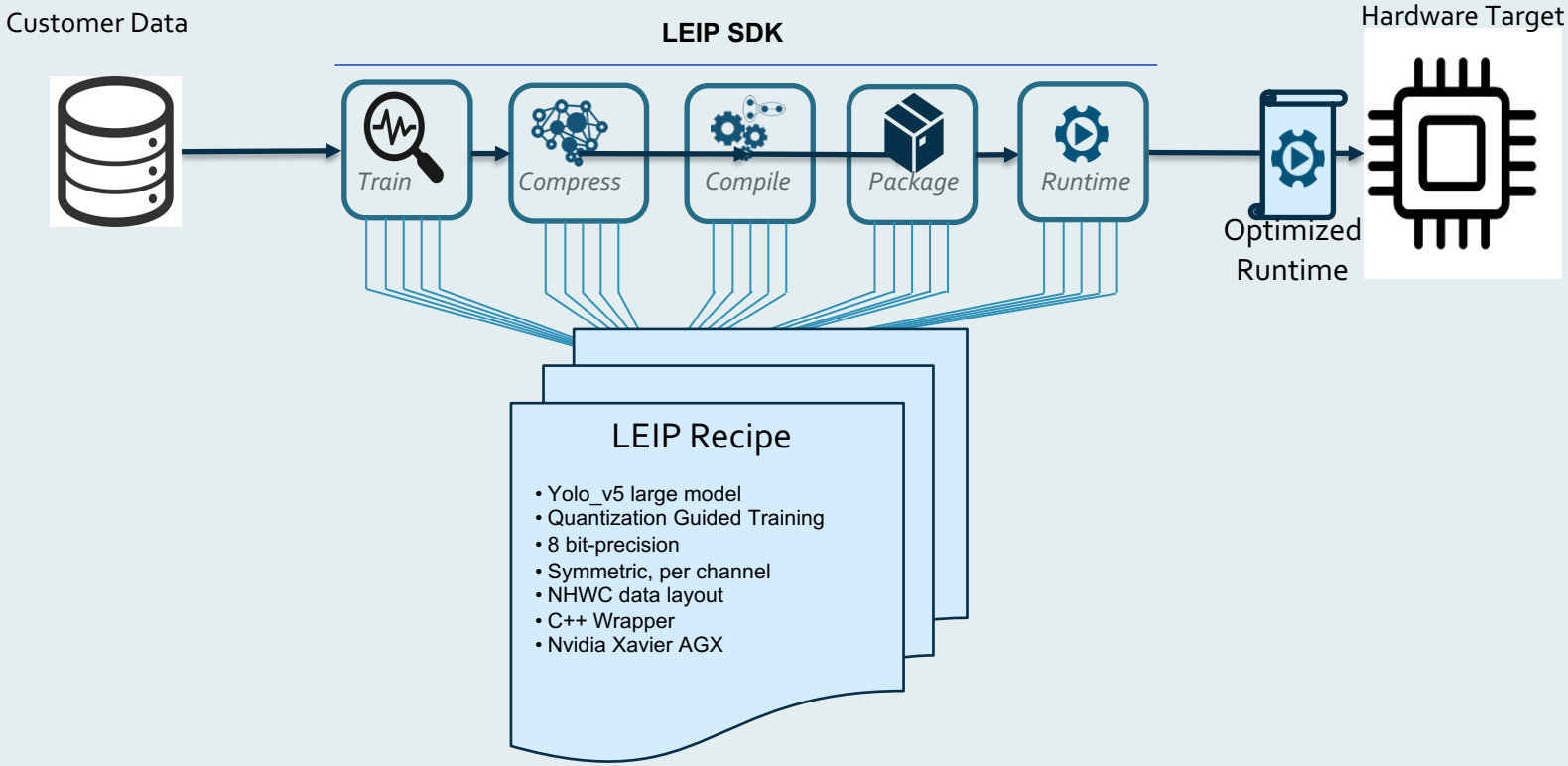
Challenges:	Resource Constraints	Developer Frustration	Difficult to Iterate
Pain points	• Inference Speed • Model Size • Memory • Cost • Power	• Hardware dependent Workflows • Manual Optimization • Poor Runtime Support	• Lack of Diagnostics • Irreproducible Models • Changing data landscape



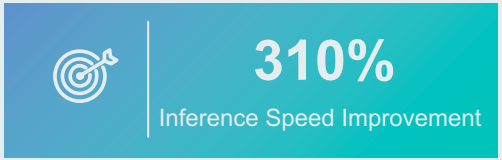
Is there an easier way to do this?

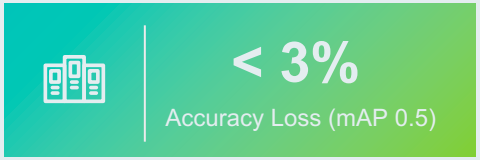
ML Engineer
Evelyn

LEIP Recipes



Recipe 1 - YoloV5 + AGX + COCO





- Pre-configured and optimized
- Demonstrates best-in-class performance
- Customers can BYOD

Example performance vs PyTorch: YOLOv5 Large, AGX, Int8 quantization

Recipe Configuration

Example: Train Yolov5 on MS COCO

```
af --config-dir=recipes --config-name=yolov5_L_RT
```

recipes/yolov5_L_RT.yaml

```yaml
defaults:
  - hydra: defaults
  - data: torchvision/coco-detection
  - model: zhiqwang-yolov5-rt-stack
  - trainer: ddp
  - callbacks: defaults
  - eval: coco
  - predict: defaults
  - vizdata: defaults
  - export: defaults
  - paths: defaults
  - command: train
  - override /hydra/job_logging: colorlog
  - _self_
task:
  width: 640
  height: 640
  batch_sizes: [8, 8]
  num_workers: 4
  moniker: recipe_${model.module.model_architecture}
```

Data

af/configs/data/torchvision/coco-detection.yaml

```yaml
nclasses: 80
module:
  _target_: af.core.data.modules.adaptermodule.AdaptorDataModule
  batch_sizes: ${task.batch_sizes}
  num_workers: ${task.num_workers}
  adaptors: ${model.adaptors}
  train_transforms:
    _target_: af.core.data.augmentations.basic.resize
    width: ${task.width}
    height: ${task.height}
  valid_transforms:
    _target_: af.core.data.augmentations.basic.resize
    width: ${task.width}
    height: ${task.height}
  dataset_generator:
      _target_: af.core.data.sets.adaptors.torchvision-coco.MSCoco
      root_path: ${paths.cache_dir}
      task_family: "detection"
```

Model

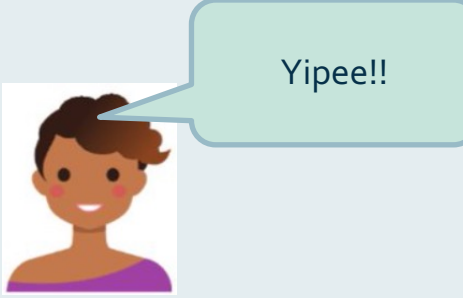
af/configs/model/zhiqwang-yolov5-rt-stack.yaml

```yaml
adaptors:
  uniform2trainvaltest:
    _target_: af.core.models.zoo.zhiqwang-yolov5-rt-stack.uniform2trainvaltest
  uniform2prediction:
    _target_: af.core.models.zoo.zhiqwang-yolov5-rt-stack.uniform2prediction
  prediction2uniform:
    _target_: af.core.models.zoo.zhiqwang-yolov5-rt-stack.prediction2uniform

trace_for_leip:
  _target_: af.core.models.zoo.zhiqwang-yolov5-rt-stack.trace_for_leip

module:
  _target_: af.core.models.zoo.zhiqwang-yolov5-rt-stack.YoloClassFactory
  patch_dir: ${paths.patch_dir}
  hash: 06022fd62dc247f1140e34bb0745673bce95ccad
  patch_file: zhiqwang-yolov5-rt-stack/20220225.patch
  model_architecture: yolov5s #yolov5s, yolov5m, yolov5l, yolov5x
  img_height: ${task.height}
  img_width: ${task.width}
  lr: 0.01
  score_thresh: 0.005 #0.005 is value tested for Coco
  num_classes: ${data.nclasses}
  pretrained: False # if you set pretrained: True, you must use the Coco dataset so num_classes match
```

Advantages of Recipes



ML
Engineer
Evelyn

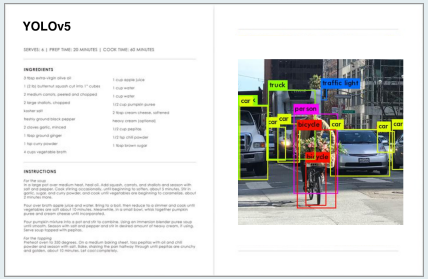
Highly Optimized Model

Exceptional Runtime Support

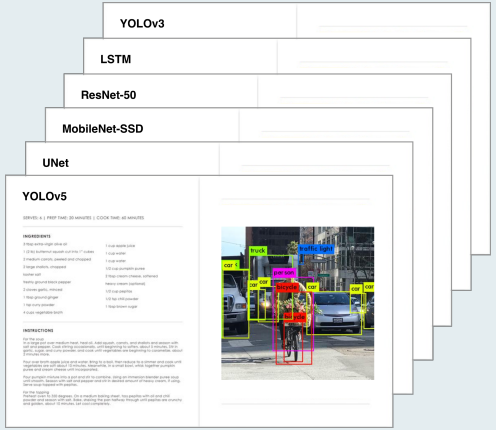
Robust, Repeatable, Reproducible Models

Recipes accelerate your time to market by 10x

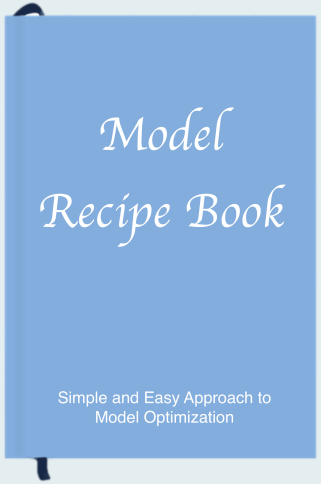
Scaling Recipes



Optimize for a single model architecture and a hardware target



Optimize for multiple model architectures running across a heterogeneous compute environment



More models and more hardware targets with more optimization techniques including throttling and pruning

Next Steps

If you want to -

- Reduce your Bill of Materials
- Fit multiple models into a single target hardware
- Accelerate your time-to-market for Edge AI Deployment

…we'd love to help! Please contact us at info@latentai.com.

Thank You

Jags Kandasamy, CEO
jags@latentai.com
+1 (404) 790-2048

Young Yoon, Products
young@latentai.com
+1 (510) 390-2041

Haya Sridharan, Products
haya@latentai.com
+1 (206) 330-7817

LatentAI

tinyML Summit 2022 Sponsors

Copyright Notice

www.tinyml.org