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
An Edge ML Story

Solutions Steve (Links with customer)

ML Engineer Evelyn

Data Scientist Daisy

Let’s detect potholes!

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

Yikes!!
Edge ML Challenges

Challenges:

Resource Constraints:
- Inference Speed
- Model Size
- Memory
- Cost
- Power

Developer Frustration:
- Hardware dependent Workflows
- Manual Optimization
- Poor Runtime Support

Difficult to Iterate:
- Lack of Diagnostics
- Irreproducible Models
- Changing data landscape

Is there an easier way to do this?

ML Engineer
Evelyn
LEIP Recipes

Customer Data → Train → Compress → Compile → Package → Runtime → Hardware Target

LEIP Recipe
- Yolo_v5 large model
- Quantization Guided Training
- 8 bit-precision
- Symmetric, per channel
- NHWC data layout
- C++ Wrapper
- Nvidia Xavier AGX

LEIP is Latent AI Efficient Inference Platform
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

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

```yaml
recipes/yolov5_L_RT.yaml
```

---

**Data**

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

```yaml
mclasses: 80
module:
  Target: af.core.data.module.AdaptorDataModule
  batch_sizes: [task.batch_sizes]
  num_workers: $task.num_workers
  adapters: $model.adapters
  train_transforms:
    _target_: af.core.data.base 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.datasets.adapters.torchvision-coco.MSCoco
    root_path: $(paths.cache_dir)
    task_family: "detection"
```

---

**Model**

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

```yaml
adapters:
  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: 4602c502c47014cbe34b6745673bce95ccad
  patch_file: zhiqwang-yolov5-rt-stack.02B2B225.patch
  model_architecture: yolov5s # yolos, yolov5m, yolov5l, yolov5x
  img_height: $task.height
  img_width: $task.width
  lr: 0.001
  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
```

---

LatentAI
Advantages of Recipes

- 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
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.tinylml.org