How To Reduce AI Bias with Synthetic Data for Edge Applications
<table>
<thead>
<tr>
<th>Date</th>
<th>Title</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 22, 2020</td>
<td>How To Reduce AI Bias with Synthetic Data for Edge Applications</td>
<td>Dori AI</td>
</tr>
<tr>
<td>October 20, 2020</td>
<td>Optimizing Power and Performance For Machine Learning at the Edge - Model Deployment Overview</td>
<td>Arm</td>
</tr>
<tr>
<td>November 3, 2020</td>
<td>Small is big: Making Deep Neural Nets faster, smaller and energy-efficient on low power hardware</td>
<td>DeepLite</td>
</tr>
</tbody>
</table>
ABOUT THE SPEAKER
Dr. Nitin Gupta, VP Product/Founder @ Dori AI
www.dori.ai

PREVIOUS ALUMNI
• Product Lead @ Google Daydream (CV/AR/VR)
• Systems Eng Lead @ Pebble/Qualcomm
• Ph.D. Advised by Steve Furber (Co-founder of ARM)

ABOUT DORI
• Full Stack Computer Vision Development Platform
• Accelerate AI+CV development for quicker time-to-market
OUTLINE

What is synthetic data? Why use it? Why now?
Data augmentation vs synthetic data?
How to leverage synthetic data in the real world?
What workflow is needed to leverage synthetic data?
How to deal with data + model bias?
How do you deploy edge applications that can leverage synthetic data?
What is synthetic data generation?

Synthetic Data Generation
Artificially generating data to meet the needs or conditions that are not available in existing real data

Two Primary Types:
- Fully Synthetic - Does not contain any real data
- Partially Synthetic - Inducing noise into the real data to simulate additional use cases
Why is it necessary?

Synthetic data fill in the gaps:

- Missing or non-existent data
- Occluded objects or scenes
- Captures different conditions
  - Camera angles / perspectives
  - Lighting
  - Environment / backgrounds
  - Motion blur
  - Pose
- Can be used to interpolate data across video frames
What are some industry applications are leveraging synthetic data?

**Manufacturing**
- product / part generation
- defects / anomalies

**Autonomous Vehicles**
- Simulating road scenarios
- Different road conditions

**Smart Cities**
- City planning
- Site surveys / reconstruction
Data Augmentation Techniques

**Color Space**
- RGB
- HSV
- YCrCb
- LAB

**Morphological**
- rotation
- translation
- flipping
- resizing

**Thresholding**
- binary
- inverse

**Filtering**
- averaging
- gaussian
- median

**Color Space Diagram**
- RGB
- YCrCb
- HSL

**Thresholding Examples**
- Original Image
- Adaptive Threshold Mean
- Adaptive Threshold Gaussian

**Filtering Techniques**
- Average Blur
- Gaussian Blur
- Median Blur
- Bilateral Filtering
Generative Adversarial Networks Techniques

Many to choose from:
- Generate New Images
- Generate Photorealistic Images
- Style Transfer
- Semantic-Image-to-Photo
- Face Generation
- Pose Generation
- Super Resolution
- Motion Prediction
Generative Adversarial Nets, Goodfellow, 2014

GP-GAN: Towards Realistic High-Resolution Image Blending, 2017

Progressive Growing of GANs for Improved Quality, Stability, and Variation, 2017

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2015

Generating Videos with Scene Dynamics, 2016

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, 2017

Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis, 2017


Pose Guided Person Image Generation, 2017
Why is synthetic data generation important now?

Industry Challenges
- Volume of data limited
- Edge/corner cases are hard to capture
- Data bias + imbalances exists in many datasets
- Access to private data is becoming harder
How do you actually leverage these data generation techniques in a real-world application?
Formula for success + velocity:
Leverage a standard workflow for all AI solutions
Problem: Most datasets are imbalanced

Sample Bias / Class Imbalance
- Sample datasets are not representative of reality
- One class too few or too many examples in the training dataset

Negative Set Bias
- Dataset does not have enough negative use cases
- Quite common in manufacturing use cases where images of defects are under represented
How do you measure data imbalances?
Analyze various metrics to determine imbalances

**Metrics**

- disparate impact
- difference in means
- difference in residuals
- normalized mutual information score
- label distribution
- statistical analysis
Solution: You need to rebalance the dataset

Most enterprises will not have all the data you need to build an accurate model. You must complement their datasets with additional data.

End Result:
- Increases model accuracy
- Improves model robustness
- Fills in missing data
- Generates negative use cases
How do you set up a proper data pipeline to generate datasets and remove data bias?

DATA PIPELINE

Original Datasets → GAN DATA GENERATION ENGINE → DATA AUGMENTATION ENGINE → DATA BIAS ANALYZER → DATASET SPLIT + MERGE ENGINE → Balanced Synthetic + Augmented Datasets

Dataset Viewer
Problem: Removing data bias does not necessarily remove model bias.

How do you ensure your model is unbiased even after training with an unbiased dataset?

Considerations:

- Data bias may or may not affect model bias
- Balancing datasets may not yield desired results - you may actually need to induce data bias
- Must look at what activations are present
- Retraining with entire rebalanced dataset is preferred rather than incremental retraining
Solution: Benchmark and analyze model bias

Model bias metrics:
- average odds difference
- disparate impact
- statistical parity difference
- gradient analysis
- pixel level feature analysis

Impact:
- Avoids overconfident or misclassifying models
- Deep understanding of what features contribute to predictions
- Obtain detailed metrics to update customize model to remove bias
What challenges do edge deployments bring to the table?

Data collection can be difficult
- Edge or on-premise environments may be inaccessible
- Data may be kept private / secure
- Synthetic data may be your only option

Be careful of model optimizations:
- Ensure edge optimizations (pruning/quantization/etc) do not introduce any biases
- Benchmark the optimized model on actual data
Once the model is deployed, how do we ensure bias or drift does not happen?

Analyze + Retrain + Redeploy

- You will not have all the data you need from the field
- You must continuously monitor your deployed models and collect runtime data for auditing
- Rebalance datasets with newly collected data to ensure robustness
Dori AI
End-to-end computer vision application development platform

❖ Connect any image / video source
❖ Augment, generate & annotate datasets
❖ Build and deploy computer vision models for any use case across edge device, edge server, or cloud
❖ Gain model and data insights via analytic dashboards
Dori Vision: A full-stack end-to-end deep learning computer vision pipeline

- COLLECT DATA
- ANNOTATE
- AUGMENT + GENERATE
- TRAIN + OPTIMIZE
- BENCHMARK + ENSEMBLE
- DEPLOY + MONITOR
- RETRAIN + REDEPLOY
DATA PIPELINE
connect + annotate + generate + augment

1. Connect + prepare image / video streams

CONNECT DATA
Format: jpg, png, bmp, mp4, avi, etc
Quality: resolution, size, frame rates
Connector Support:
- streaming
- local upload
- cloud storage
- edge devices / cameras

2. Annotate images / videos

ANNOTATE
Considerations:
- Background noise / objects
- Occlusions
- Camera angles / perspective / distance
- Lighting / blur / resolution

3. Augment existing data + generate synthetic data

SYNTHETIC DATA + DATA AUGMENTATION
Types:
- Fully synthetic
- partially synthetic
- GAN
- CV transformations
MODEL PIPELINE

there is a lot more to consider than just training

1. Select model for use case
2. Train custom model using use case specific datasets
3. Validate accuracy
4. Optimize a model for deployment
5. Benchmark model to ensure accuracy + latency on deployment HW
6. Ensemble multiple models if required for the use case
**MODEL SELECTION**

Types:
- Classification, Detection, Segmentation, Actions, Pose

Considerations:
- Pretrained models
- Model classes
- Image / video preprocessing
- Post processing logic
- Action recognition vs motion tracking?

---

**TRAIN + VALIDATE**

Considerations:
- Transfer Learning vs AutoML vs Fully Custom
- Don’t forget about production - is the model deployable?
- Don’t forget the cost of training
  - i.e. high-end GPU cloud instances can make or break the budget
- Hyperparameter tuning

---

**OPTIMIZE + BENCHMARK**

Considerations:
- Trade Offs: latency vs size vs accuracy vs cost
- Model Optimization: quantization, pruning
- System Optimization
  - HW vendor-specific
- Retraining required after optimization?
- Must benchmark on multiple datasets
- Must benchmark on deployment hardware

---

**ENSEMBLE**

Considerations:
- Solutions may require multiple models to satisfy business use cases
  - i.e. moving violation = person detection + motion tracking + boundary detection
- Latency vs throughput vs cost
PRODUCTION PIPELINE

deploy + predict + monitor + analyze + retrain + redeploy

1. Deploy models across cloud, hybrid, or edge use cases
2. Run inference
3. Feed prediction results to application / business logic
4. Collect runtime data & system metrics
5. Analyze runtime prediction results
6. Re-annotate, retrain & re-deploy models
DEPLOY
Considerations:
- Cloud
  - Scalability: Docker / Kubernetes
  - Cost vs QoS vs Customer Experience
- Edge
  - Device + model management
  - Multiple data streams

APPLICATION INTEGRATION
Considerations:
- How to consume prediction results?
  - Realtime vs offline
- How to store prediction results?
  - Local database vs cloud database
- How will results + incoming media be visualized?
  - BI Tool (i.e. Tableau) vs custom dashboard

MONITOR
Considerations:
- Collect data + statistics
- Image + video data sampling
- Prediction results
- System performance metrics
- Multiple camera streams

ANALYZE + RETRAIN + REDEPLOY
Considerations:
- Model / data drift
- Bias - model, region, specific deployments
- Anomalies / degradation
- Explainability
- Active learning loops
Thank You
Danke
Merci
謝謝
ありがとう
Gracias
Kiitos
감사합니다
धन्यवाद
شكرًا
תודה
Join us at Arm DevSummit

Oct 6 - 8 | Virtual Conference
Register here [https://devsummit.arm.com/arm-ai-ml](https://devsummit.arm.com/arm-ai-ml)