September 2, 2020

@qualcomm_tech

Pushing the boundaries of AI research

Qualcomm Technologies, Inc.

Our AI leadership

Over a decade of cutting-edge AI R&D, speeding up commercialization and enabling scale

2007
Qualcomm Research initiates first AI project

2009
- Research in spiking neural networks
- Investment and collaboration with Brain Corp
- Deep-learning based AlexNet wins ImageNet competition

2013
- Research artificial neural processing architectures
- Research face detection with deep learning
- Completed Brain Corp joint research

2015
- Opened Qualcomm Research Netherlands
- Acquired EuVision
- MWC demo showcasing photo sorting and handwriting recognition
- Opened joint research lab with University of Amsterdam
- 1st Gen Qualcomm® AI Engine (Qualcomm® Snapdragon™ 820 Mobile Platform)

2016
- Collaboration with Google on TensorFlow
- Qualcomm® Neural Processing SDK
- 2nd Gen Qualcomm AI Engine (Snapdragon 835)

2017
- Announced Facebook Caffe2 support
- Brain Corp raises $114M
- Snapdragon 660 Snapdragon 630
- Acquired Scyfer
- 3rd Gen Qualcomm AI Engine (Snapdragon 845)

2018
- Qualcomm Technologies ships ONNX supported by Microsoft, Facebook, Amazon
- Qualcomm Technologies researchers win best paper at ICLR
- Qualcomm® Artificial Intelligence Research initiated
- Qualcomm® Vision Intelligence Platform
- Snapdragon 710

2019
- 3rd Gen Snapdragon Automotive Cockpit
- 4th Gen Qualcomm AI Engine (Snapdragon 855)
- Power efficiency gains through compression, quantization, and compilation
- Gauge equivariant CNNs
- Qualcomm® QCS400 (First audio SoC)
- Qualcomm® Cloud AI 100
- Snapdragon 665, 730, 730G

2020
- Qualcomm® Snapdragon Ride™ Platform
- 5th Gen Qualcomm AI Engine (Snapdragon 865)
- Mobile AI Enablement Center in Taiwan to open
- AI Model Efficiency Toolkit (AIMET) open sourced
- Qualcomm® Robotics RB5 Platform

Consistent AI R&D

investment is the foundation for product leadership

Qualcomm Artificial Intelligence Research is an initiative of Qualcomm Technologies, Inc. Qualcomm Snapdragon, Qualcomm Neural Processing SDK, Qualcomm Vision Intelligence Platform, Qualcomm AI Engine, Qualcomm Cloud AI, Qualcomm Snapdragon Ride, Qualcomm Robotics RB3 Platform, and Qualcomm QCS400I are products of Qualcomm Technologies, Inc. and/or its subsidiaries. AI Model Efficiency Toolkit is a product of Qualcomm Innovation Center, Inc.

Advancing AI research to make efficient AI ubiquitous

Power efficiency
Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization
Continuous learning, contextual, always-on, privacy-preserved, distributed learning

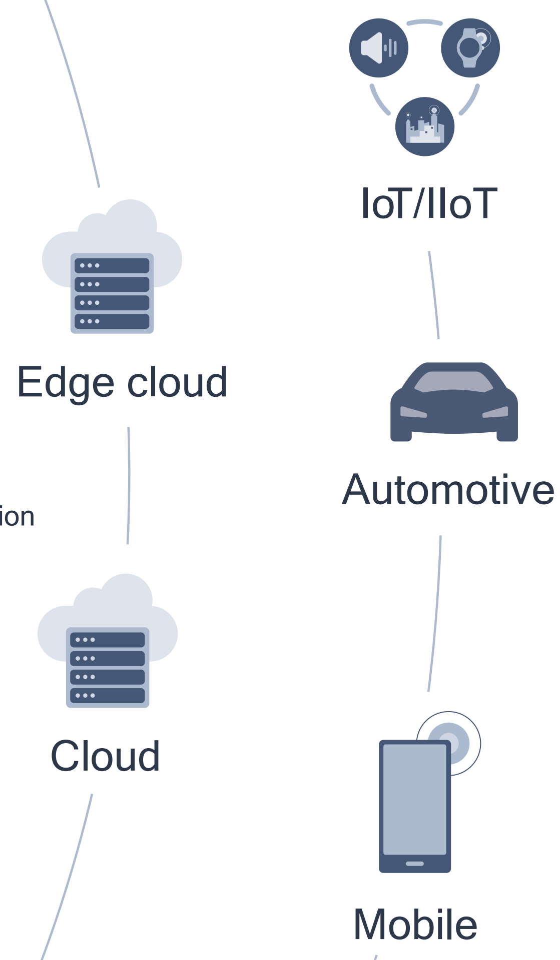
Efficient learning
Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception
Object detection, speech recognition, contextual fusion

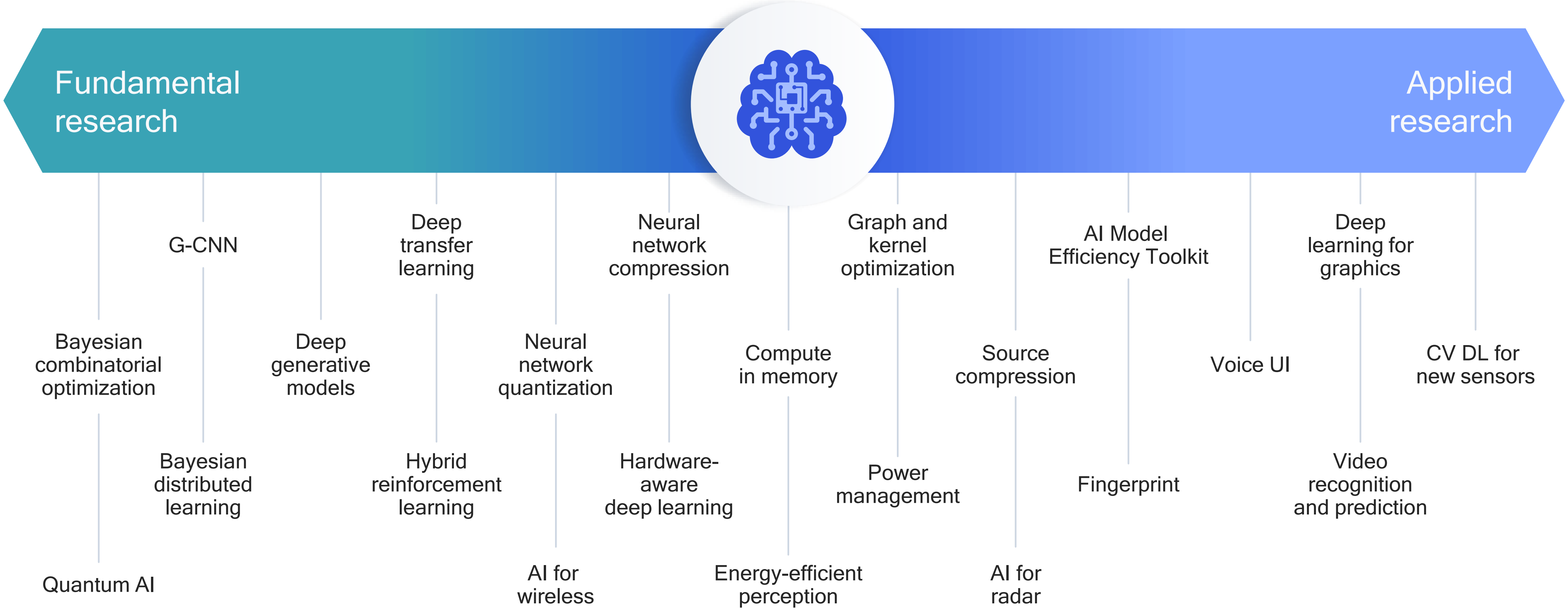
Reasoning
Scene understanding, language understanding, behavior prediction

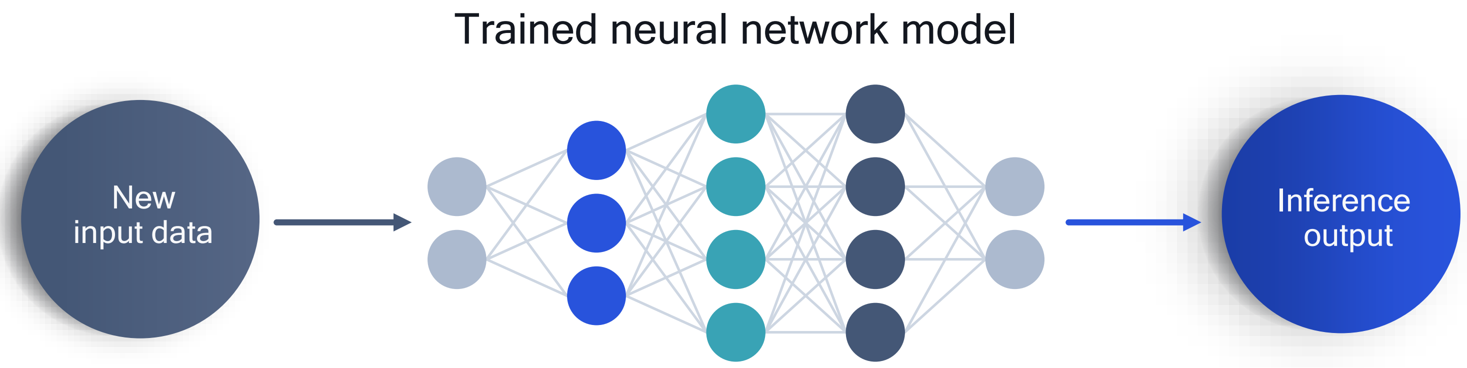
Action
Reinforcement learning for decision making



Leading research and development

Across the entire spectrum of AI



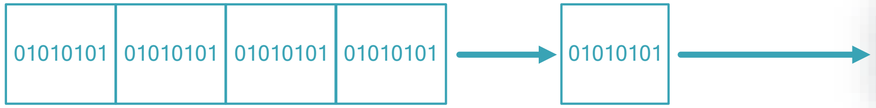


Quantization

Automated reduction in precision of weights and activations while maintaining accuracy

Models trained at high precision

Inference at lower precision



32-bit Floating point
3452.3194

8-bit Integer
3452

>4x

Increase in performance per watt from savings in memory and compute[1]

Promising results show that low-precision integer inference can become widespread

Virtually the same accuracy between a FP32 and quantized AI model through:

- Automated, data free, post-training methods
- Automated training-based mixed-precision method

1: FP32 model compared to a INT8 quantized model

Leading research to efficiently quantize AI models

Pushing the limits of what's possible with quantization

Data-free quantization

How can we make quantization
as simple as possible?

Created an automated method that addresses bias and imbalance in weight ranges:

- ☑ No training
- ☑ Data free

AdaRound

Is rounding to the nearest value the best approach for quantization?

Created an automated method for finding the best rounding choice:

- ☑ No training
- ☑ Minimal unlabeled data

Bayesian bits

Can we quantize layers to different bit widths based on precision sensitivity?

Created a novel method to learn mixed-precision quantization:

- ☑ Training required
- ☑ Training data required
- ☑ Jointly learns bit-width precision and pruning

SOTA 8-bit results

Making 8-bit weight quantization ubiquitous

>4x Increase in performance per watt while **only losing 0.5% of accuracy** against FP32 MobileNet V2

Data-Free Quantization Through Weight Equalization and Bias Correction (Nagel, van Baalen, et al., ICCV 2019)

SOTA 4-bit weight results

Making 4-bit weight quantization ubiquitous

>8x Increase in performance per watt while **only losing 2.5% of accuracy** against FP32 MobileNet V2

Up or Down? Adaptive Rounding for Post-Training Quantization (Nagel, Amjad, et al., ICML 2020)

SOTA mixed-precision results

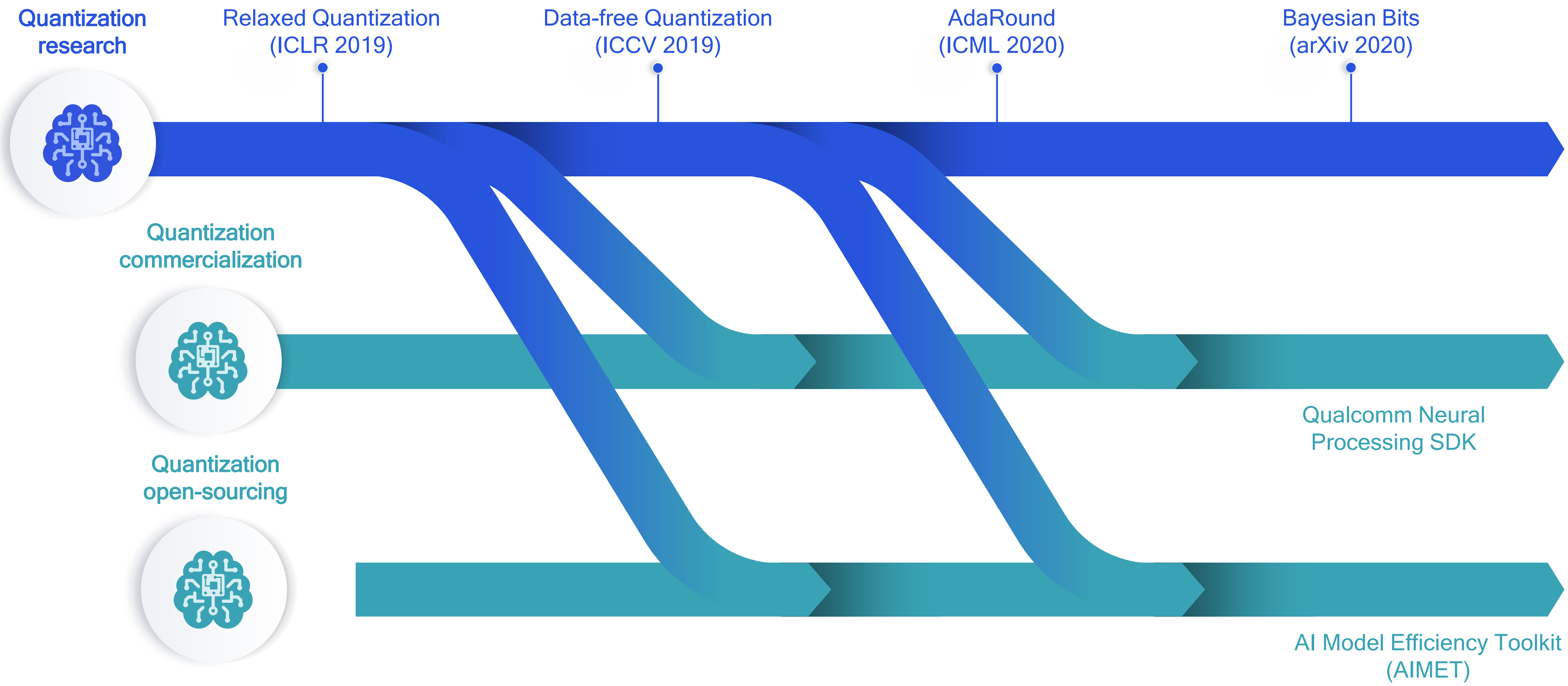
Automating mixed-precision quantization and enabling the tradeoff between accuracy and kernel bit-width

>8x Increase in performance per watt while **only losing 0.8% of accuracy** against FP32 MobileNet V2

Bayesian Bits: Unifying Quantization and Pruning van Baalen, Louizos, et al., arXiv 2020)

Leading quantization research and fast commercialization

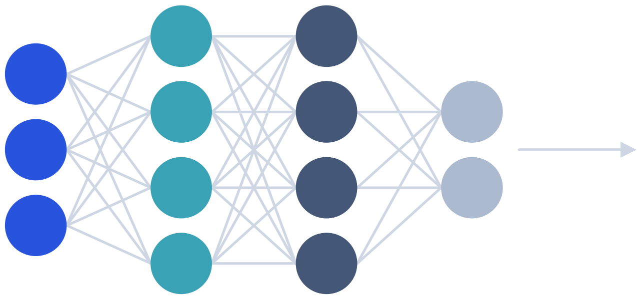
Driving the industry towards integer inference and power-efficient AI



AIMET makes AI models small

Open-sourced GitHub project that includes state-of-the-art quantization and compression techniques from Qualcomm AI Research

Trained
AI model

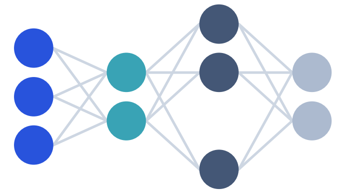


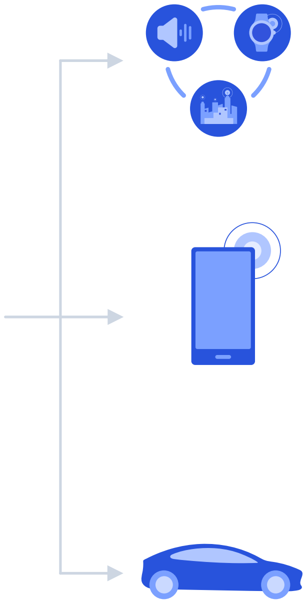
TensorFlow or PyTorch

AI Model Efficiency Toolkit
(AIMET)

- Compression
- Quantization

Optimized
AI model





If interested, please join the AIMET GitHub project: https://github.com/quic/aimet

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. AI Model Efficiency Toolkit is a product of Qualcomm Innovation Center, Inc.

Features:

- State-of-the-art network compression tools
- State-of-the-art quantization tools
- Support for both TensorFlow and PyTorch
- Benchmarks and tests for many models
- Developed by professional software developers

An increasing demand for energy efficient video processing

Valuable information is being extracted from video streams across diverse devices and use cases

Enhanced perception through object detection and semantic segmentation

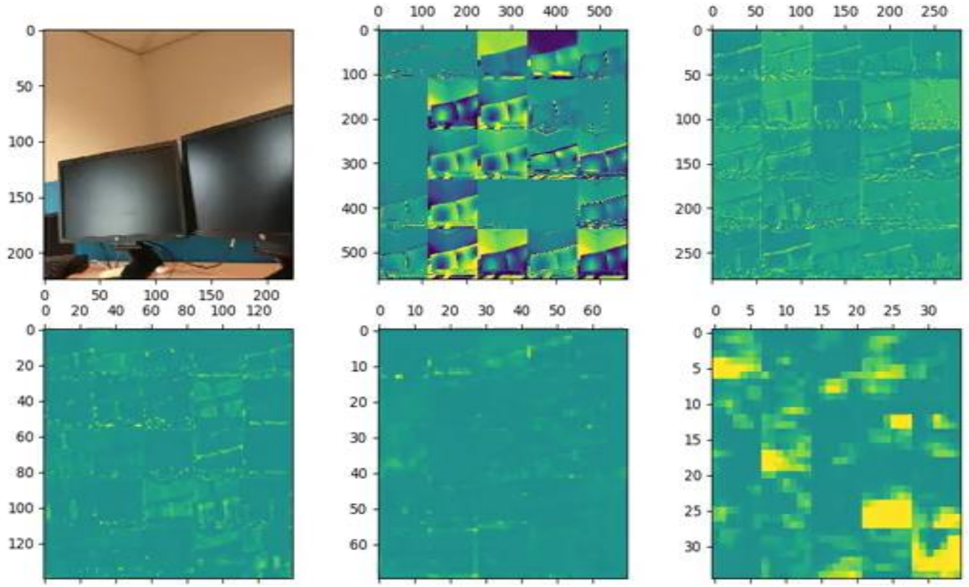
Advanced camera features, like video enhancement and super-resolution

Advanced video understanding, like search and surveillance

Increased video compression to address the demand for rich media

Developing AI approaches to reduce unnecessary computation

Recognizing redundancy between video frames so that the same thing is never computed twice



Feature maps over time for ResNet18 remain mostly constant

Tremendous redundancy across frames!

Conditional computation using gated networks

Introduce gates with low computation cost to skip unnecessary computation in neural networks

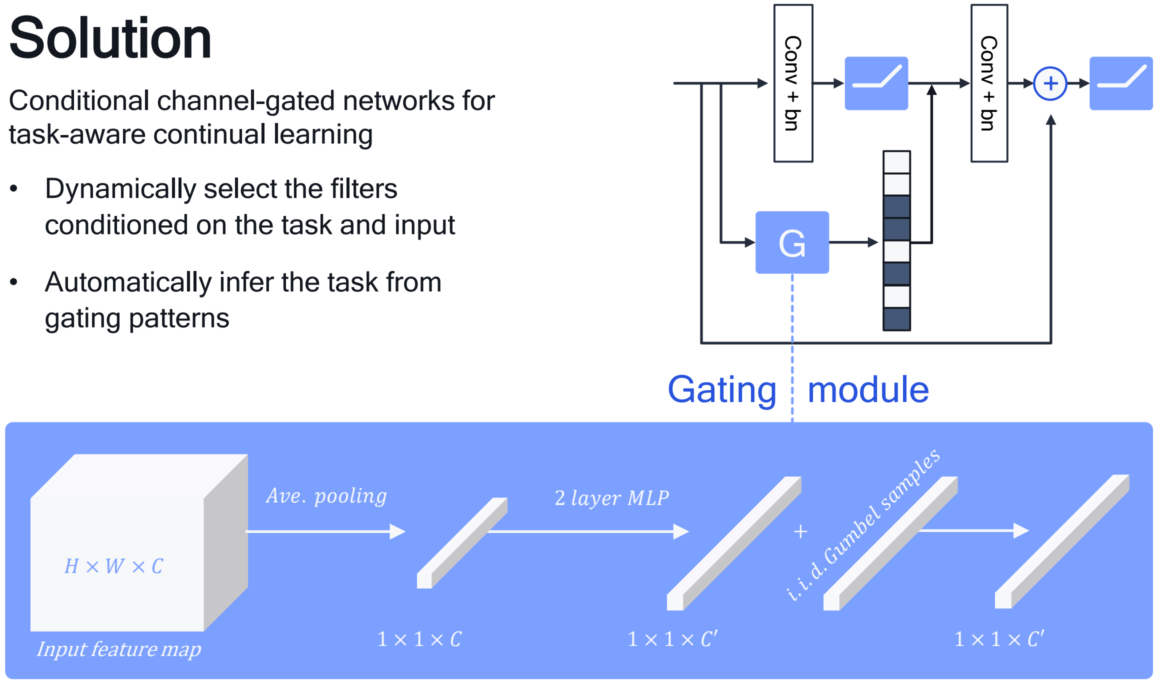
Problem

A large portion of the neural network is not necessary for the prediction, wasting computations

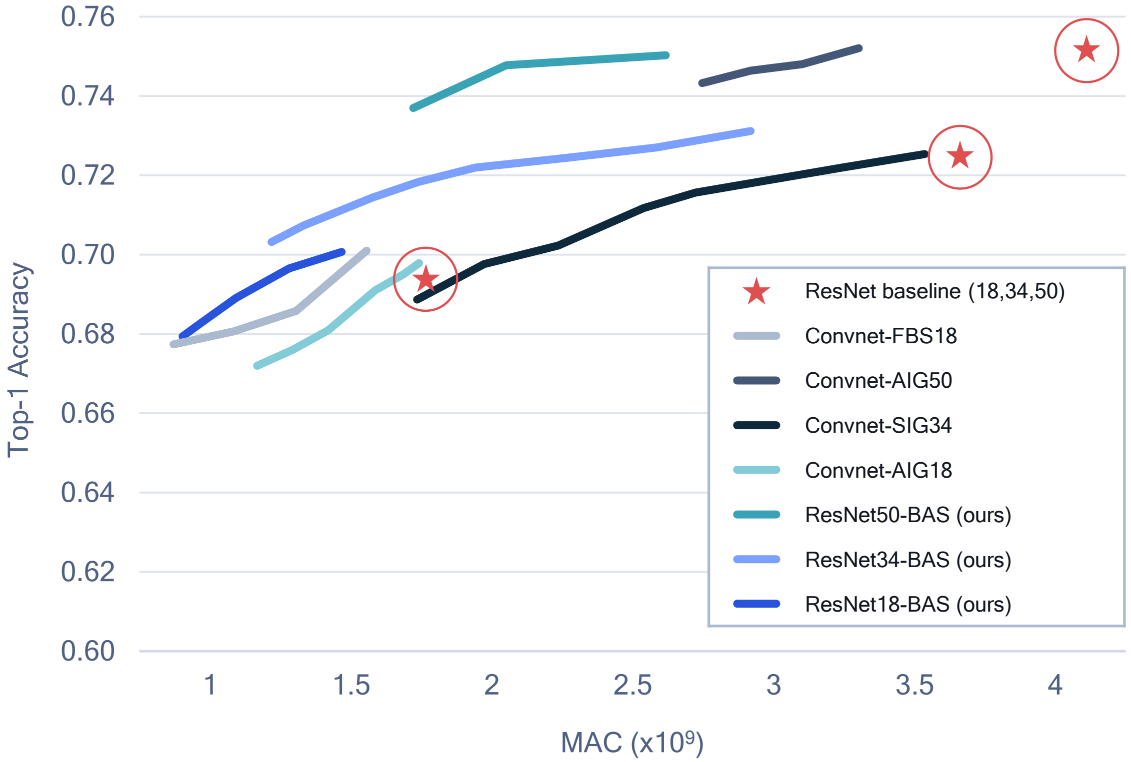
Solution

Conditional channel-gated networks for task-aware continual learning

- Dynamically select the filters conditioned on the task and input
- Automatically infer the task from gating patterns



State-of-the-art accuracy while reducing computation by up to 3X



1. Bejnordi, Babak Ehteshami, Tijmen Blankevoort, and Max Welling. "Batch shaping for learning conditional channel gated networks." ICLR (2020).
2. Abati, Davide, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. "Conditional Channel Gated Networks for Task-Aware Continual Learning." CVPR (2020).

Deep generative model research for unsupervised learning

Given unlabeled training data, generate new samples from the same distribution

Generative models

- Variational auto encoder (VAE)*
- Generative adversarial network (GAN)
- Auto-regressive
- Invertible

Powerful capabilities

- Extract features by learning a low-dimension feature representation
- Sampling to generate, restore, predict, or compress data

Broad applications

- Speech/video compression
- Text to speech
- Graphics rendering
- Computational photography
- Voice UI

* VAE first introduced by D. Kingma and M. Welling in 2013

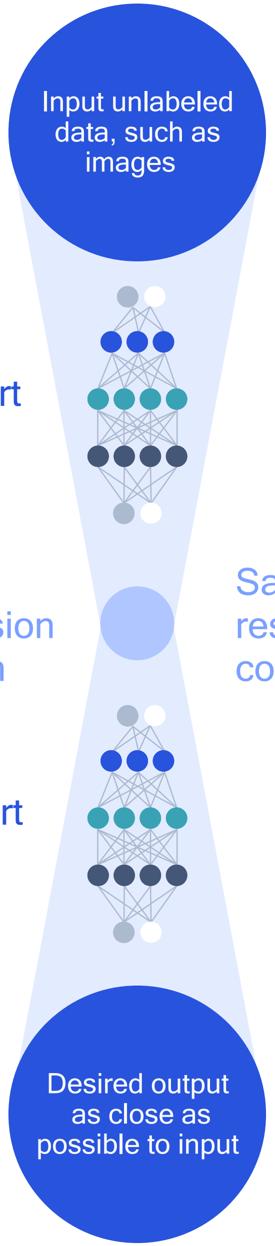
Deep generative model research for unsupervised learning

Given unlabeled training data, generate new samples from the same distribution

Generative models

- Variational auto encoder (VAE)*
- Generative adversarial network (GAN)
- Auto-regressive
- Invertible

Example use case: encoder/decoder



Encoder part

Extract features by learning a low-dimension feature representation

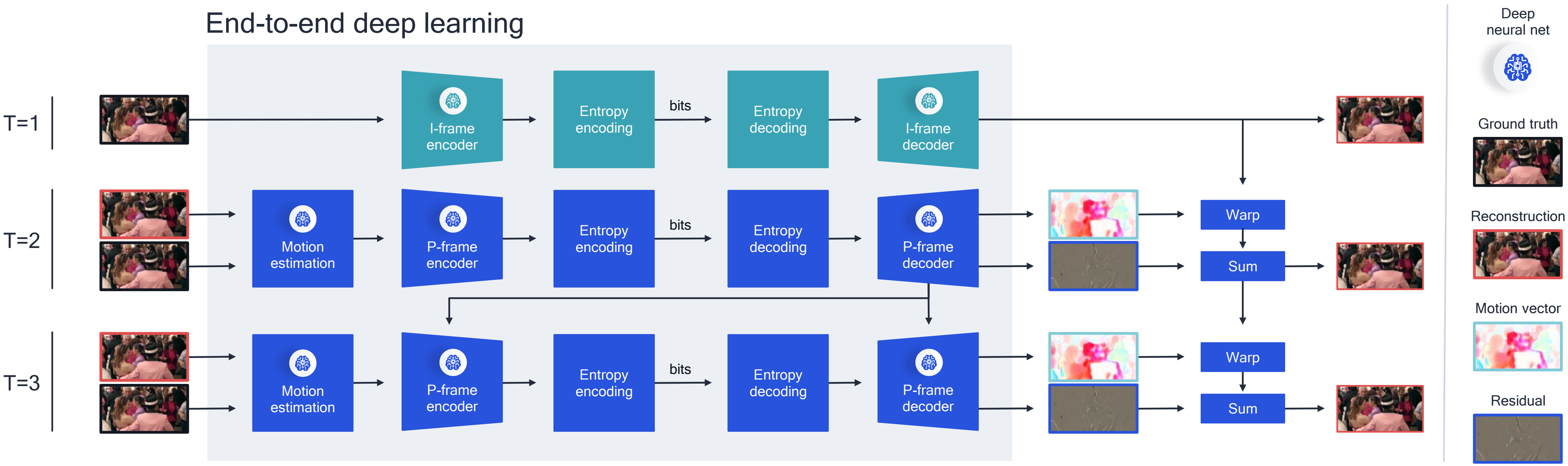
Sampling to generate, restore, predict or compress data

Decoder part

* VAE first introduced by D. Kingma and M. Welling in 2013

Novel machine learning-based video codec research

Neural network based I-frame and P-frame compression



Feedback Recurrent Autoencoder for Video Compression (Golinski, et al., arXiv 2020)
Video used in images is produced by Netflix, with CC BY-NC-ND 4.0 license: https://media.xiph.org/video/derf/ElFuente/Netflix_Tango_Copyright.txt

Achieving state-of-the-art rate-distortion compared with other learned video compression solutions

Applying AI to solve difficult wireless challenges

Deep wireless domain knowledge is required to optimally use AI capabilities

Wireless challenges

Hard-to-model problems

Computational infeasibility of optimal solution

Efficient modem parameter optimization

AI-enhanced wireless communications

AI strengths

Learning representations for hard-to-model problems

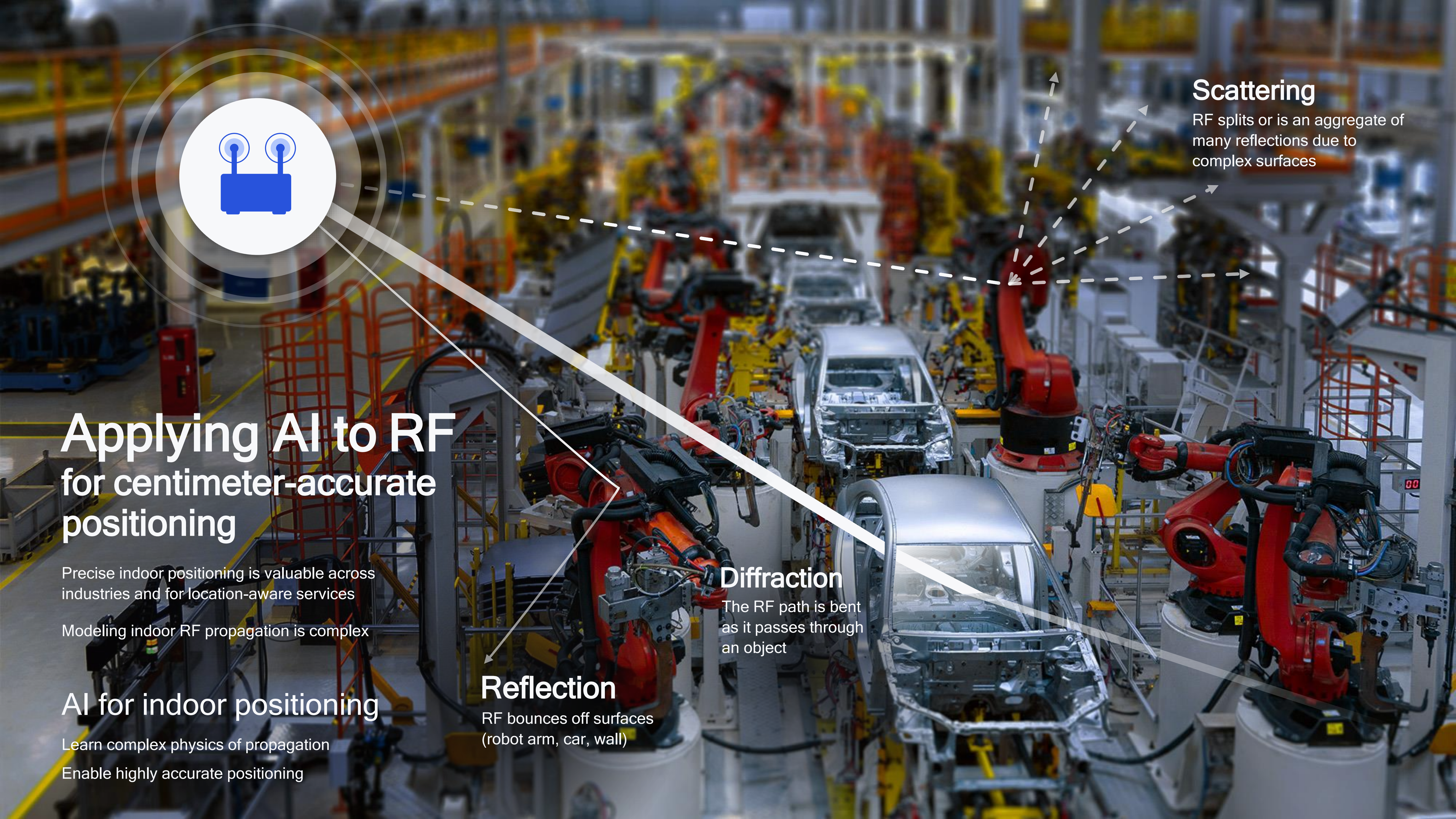
Training policies and computationally realizable solutions

Learning to compensate for non-linearities

Applying AI to RF
for centimeter-accurate positioning

Precise indoor positioning is valuable across industries and for location-aware services

Modeling indoor RF propagation is complex



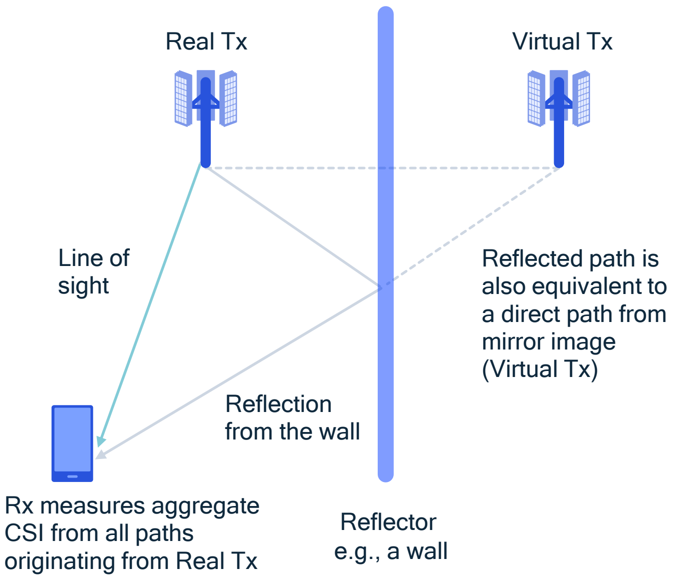
AI for indoor positioning

Learn complex physics of propagation

Enable highly accurate positioning

Neural unsupervised learning from RF for positioning

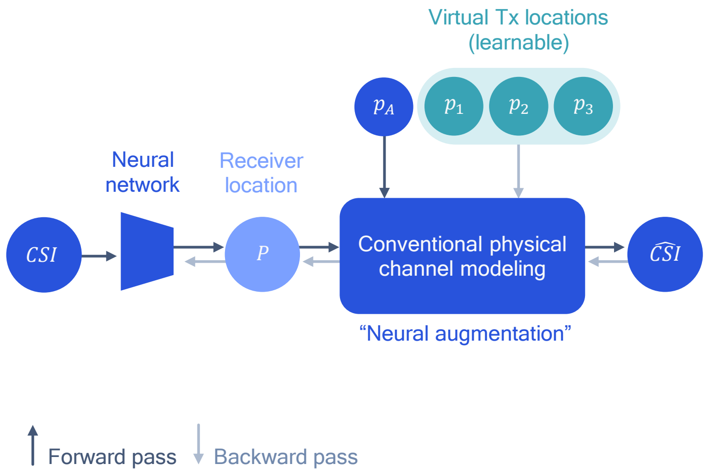
Injecting domain knowledge for interpretability



Physics of reflections

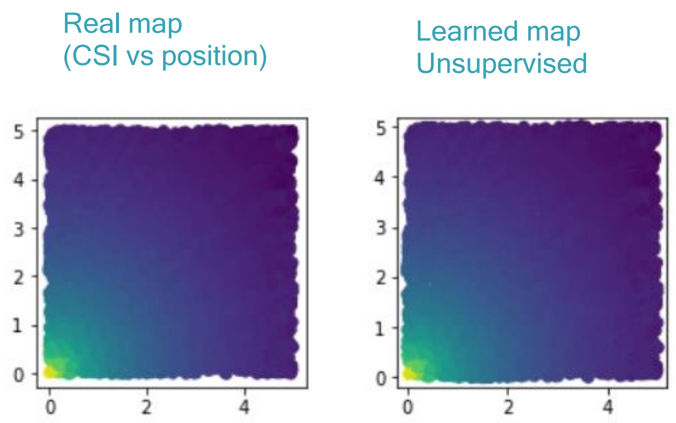
The receiver (Rx) collects unlabeled channel state info (CSI) observations throughout the environment.

The goal is to learn the Virtual Tx locations and how to triangulate using CSI.



Neural augmentation

The neural network uses a generative auto-encoder plus conventional channel modeling (based on physics of propagation) to train on the observations and learn the environment.



Incredible results

The neural network learns the virtual transmitter locations up to isometries completely unsupervised.

With a few labeled measurements, map ambiguity is resolved to achieve cm-level positioning.

Teaching Cars to See

AI RADAR

Perceptive radar

Traditional radar is affordable and responsive, has a long range, measures velocity directly, and isn't compromised by lighting or weather conditions

Applying deep learning directly to the radar signal improves virtually all existing radar capabilities

Complementary sensor

Each sensor has its own strengths and complements other sensors

A car can see best when utilizing all of its sensors together, otherwise known as sensor fusion

Research results

Significant improvements in position and size estimation, velocity estimation, object classification, and uncertainty estimation

Robust performance even in complex scenarios

Future research areas

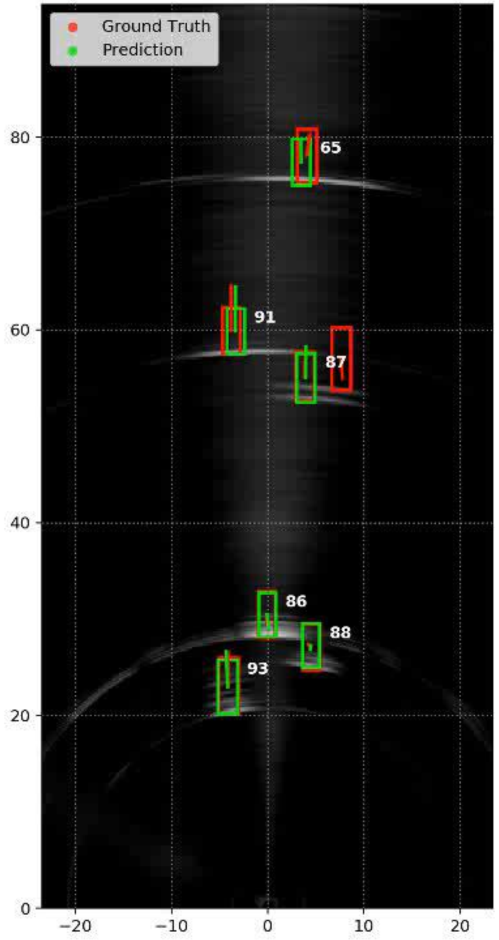
Radar compression, elevation estimation, drivable space, sparse radar sensing, pedestrian sensing, range extension, and adaptive sampling research

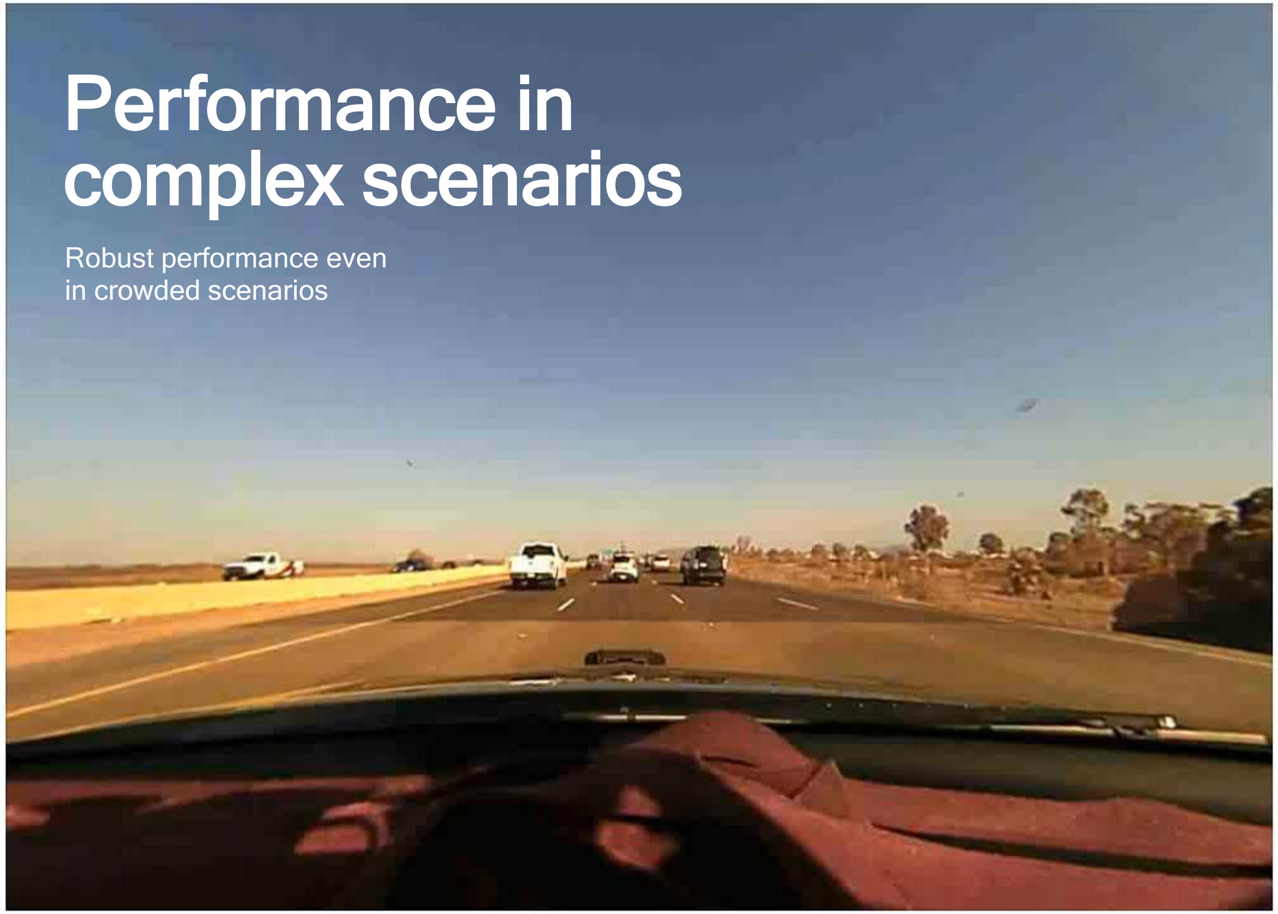
Sensor fusion research

Paving the road to autonomous driving with more perceptive radar

AI radar detects occluded vehicles in complex scenarios

Video shows the accuracy between AI radar and ground truth up to 94 meters





From Bayesian bits to quantum bits

Through quantum computing, utilize quantum mechanics to achieve exponential speedup

Superposition

Each qubit is both 1 and 0 at the same time

Entanglement

Qubits that are inextricably linked such that whatever happens to one immediately affects the other

Quantum computing

Utilize quantum mechanics to achieve exponential speedup

Classical bit

0 1

Bayesian bit

0 1

Quantum bit (qubit)

Applying quantum mechanics to machine learning

Fundamental green field research to utilize the exponential power of quantum computing on various use cases

Quantum annealing

Combinatorial optimization problems are widespread, including chip physical design and architecture search

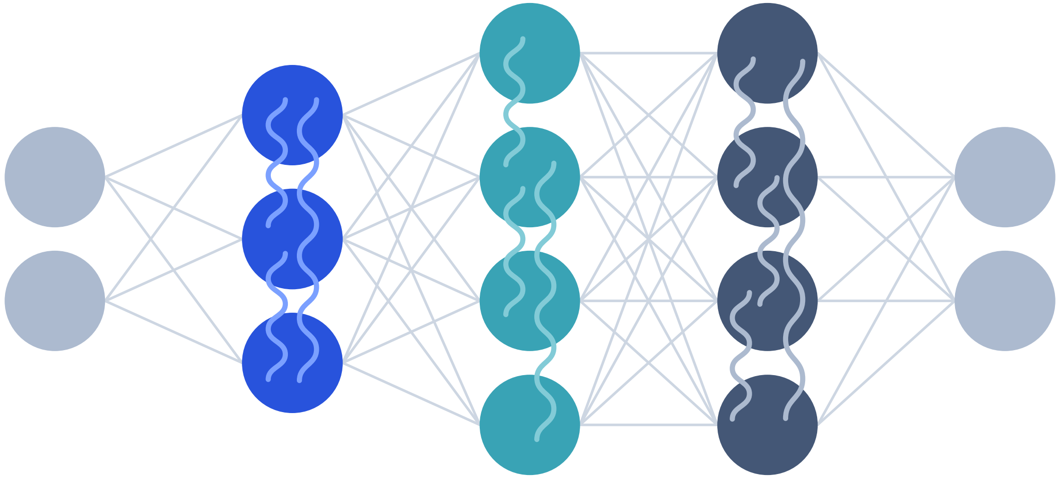
Classical computing hits its limits for a large number of states

Quantum mechanics gives fast search solutions to combinatorial problems

Quantum deep learning

The statistics of quantum mechanics can apply to deep learning

Exploration of quantum-inspired classical algorithm



Quantum binary networks are efficient on classical devices

Quantum deformed binary neural networks

Running a classical neural network on quantum computer

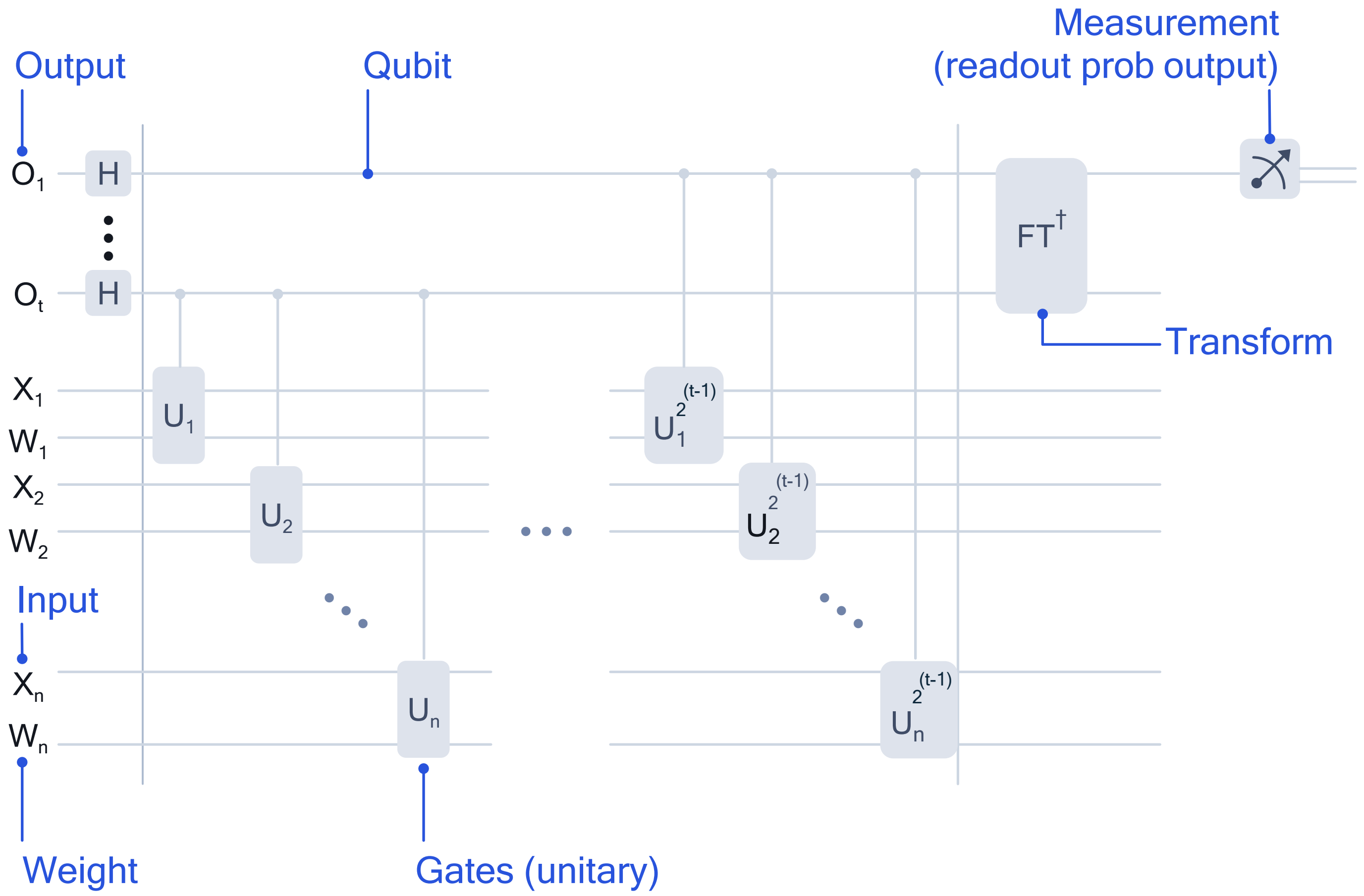
Key ideas

Deform a classical neural network into a quantum neural network with quantum gates

Run on quantum computer or efficiently simulate on classical computer

Initial result

98.7% accuracy on MNIST



First quantum binary neural network for real data!

Qualcomm

Advancing AI research to make power efficient AI ubiquitous – from device to cloud

Conducting leading research and development across the entire spectrum of AI

Creating an AI platform fundamental to scaling AI across the industry and applications

Questions?

Connect with Us

www.qualcomm.com/ai

www.qualcomm.com/news/onq

@qualcomm_tech

http://www.youtube.com/playlist?list=PL8AD95E4F585237C1&feature=plcp

http://www.slideshare.net/qualcommwirelessevolution

Thank you

Follow us on:

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

References in this presentation to "Qualcomm" may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes Qualcomm's licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm's engineering, research and development functions, and substantially all of its product and services businesses, including its semiconductor business, QCT.