July 2018

@qualcomm_tech

Qualcomm

Accelerating algorithmic and hardware advancements for power efficient on-device AI

Qualcomm Technologies, Inc.

Computers are consuming an increasing amount of energy

By 2025, the data center sector could be using 20% of all available electricity in the world[1]

A cloud provider used the equivalent energy consumption of ~366,000 US households in 2014[2]

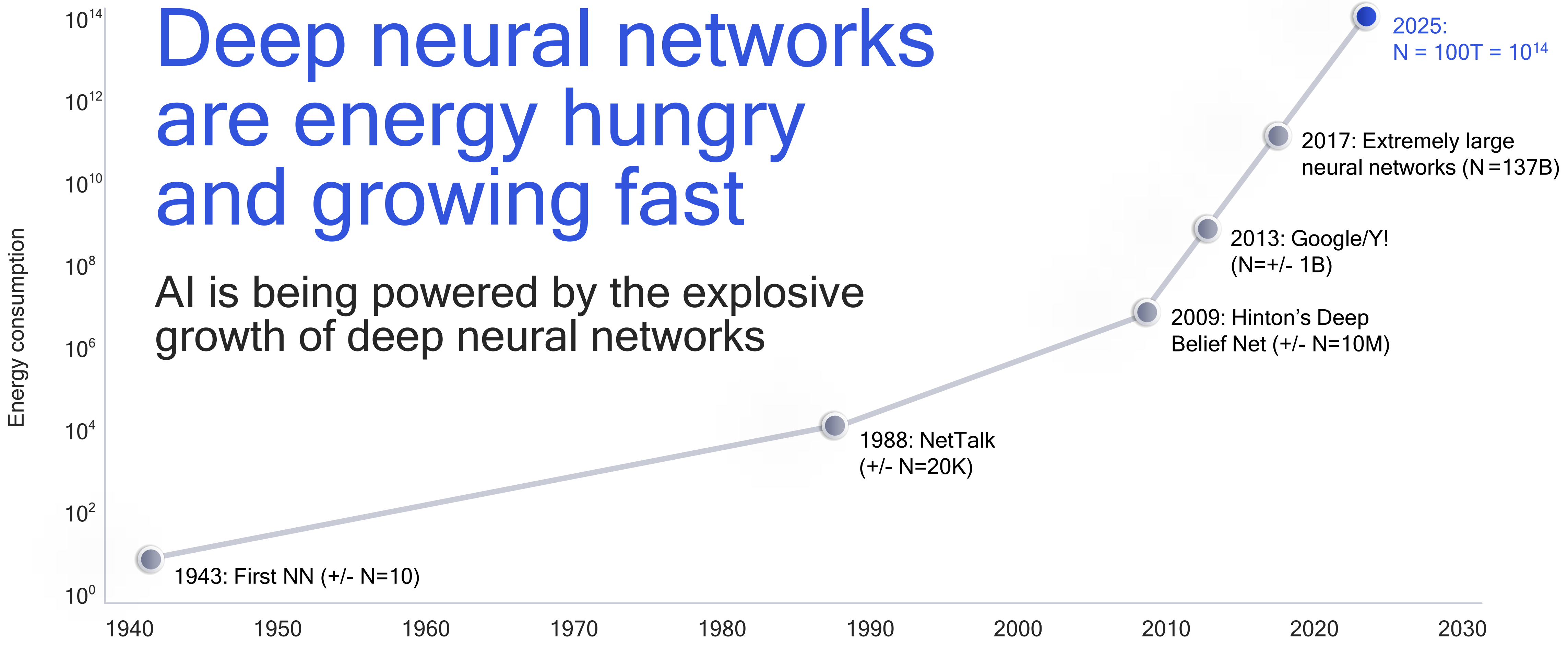
Bitcoin mining in 2017 used the same energy as did all of Ireland[3]

Given the economic potential of AI, these numbers will only be increasing

1. Andrae, Anders (2017) Total Consumer Power Consumption Forecast; 2. The Verge (2014); 3.The Guardian (2017)

Deep neural networks are energy hungry and growing fast

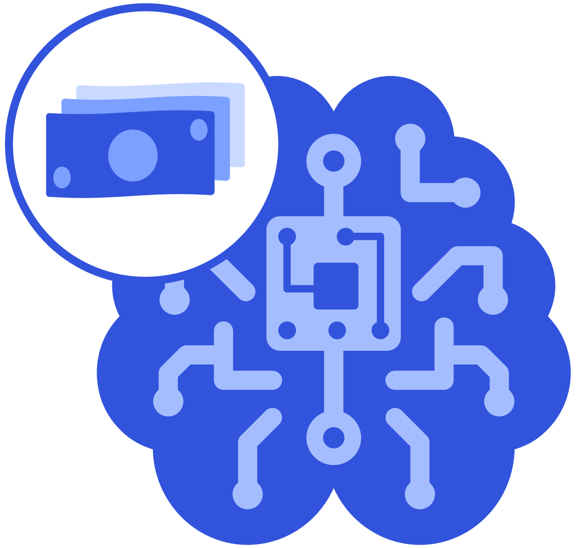
AI is being powered by the explosive growth of deep neural networks



Source: Welling

2025 | Will we have reached the capacity of the human brain?
Energy efficiency of a brain is 100x better than current hardware

Value created by AI must exceed the cost to run the service



Economic feasibility per transaction may require cost as low as a micro-dollar (1/10,000th of a cent)

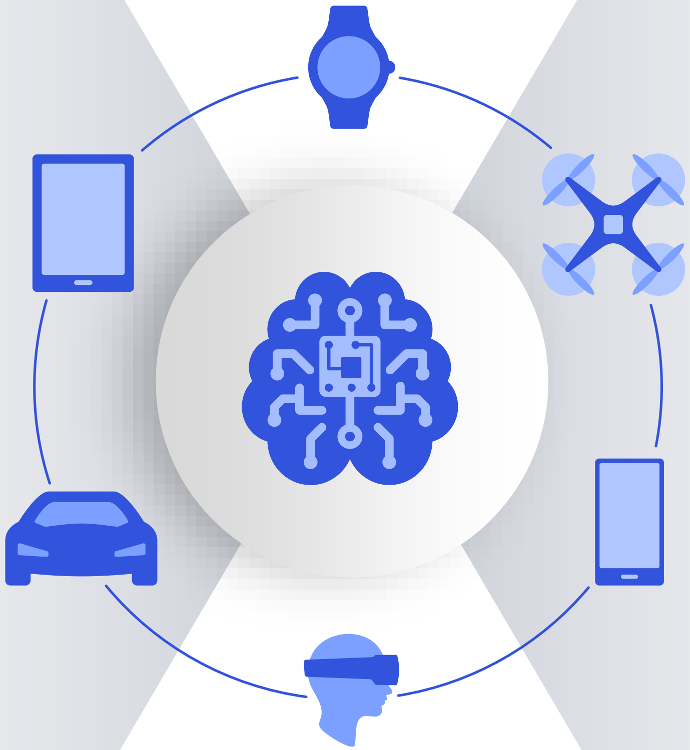
- Personalized advertisements and recommendations
- Smart security monitoring based on image and sound recognition
- Efficiency improvements for smart cities and factories

Broad economic viability requires energy efficient AI

The AI power and thermal ceiling

The challenge of AI workloads

- Very compute intensive
- Complex concurrencies
- Real-time
- Always-on



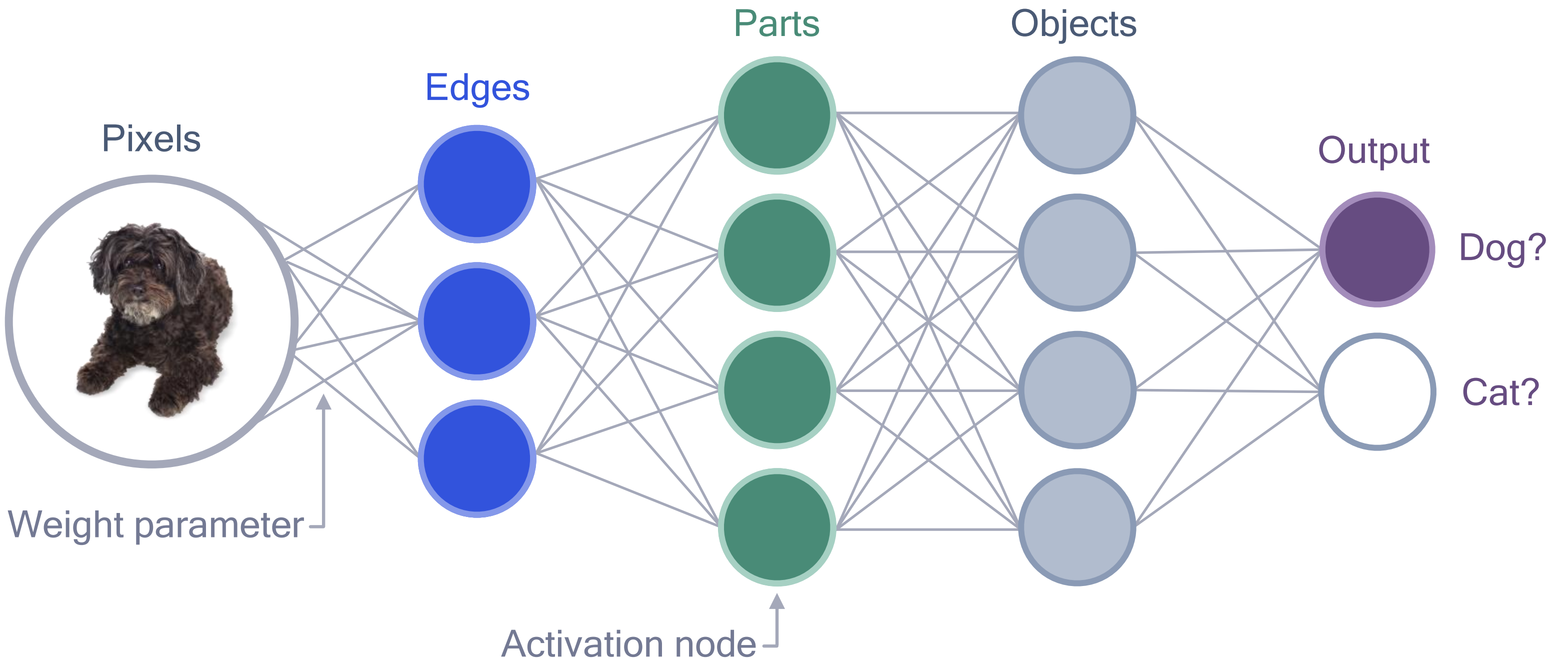
Constrained mobile environment

- Must be thermally efficient for sleek, ultra-light designs
- Requires long battery life for all-day use
- Storage/memory bandwidth limitations

Soon, AI algorithms will be measured by the amount of intelligence they provide per watt hour.

Deep learning

The good

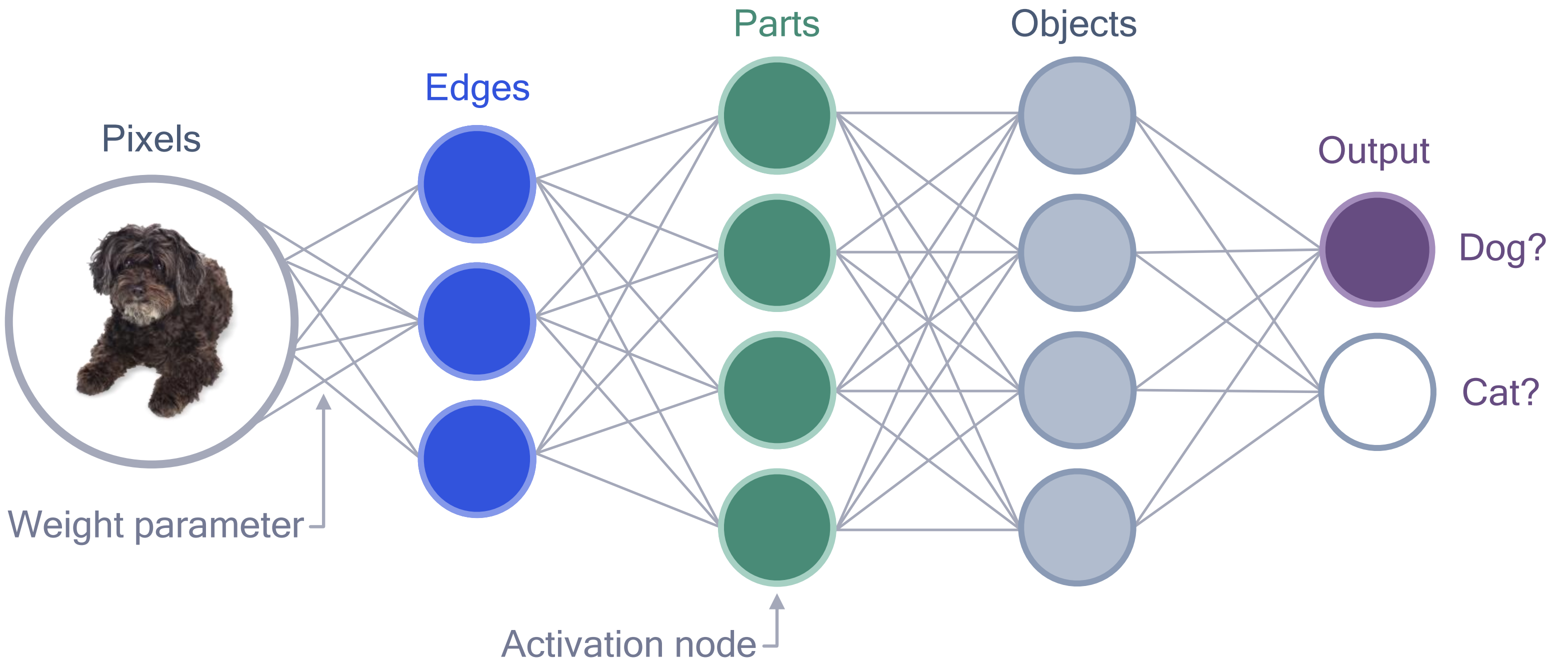


Convolutional neural networks (CNNs) have been very successful

- Extract learnable features with state-of-the art results
- Encode location invariance, namely that the same object may appear anywhere in the image
- Share parameters, making them "data efficient"
- Execute quickly on modern hardware with parallel processing

Deep learning

The bad and the ugly



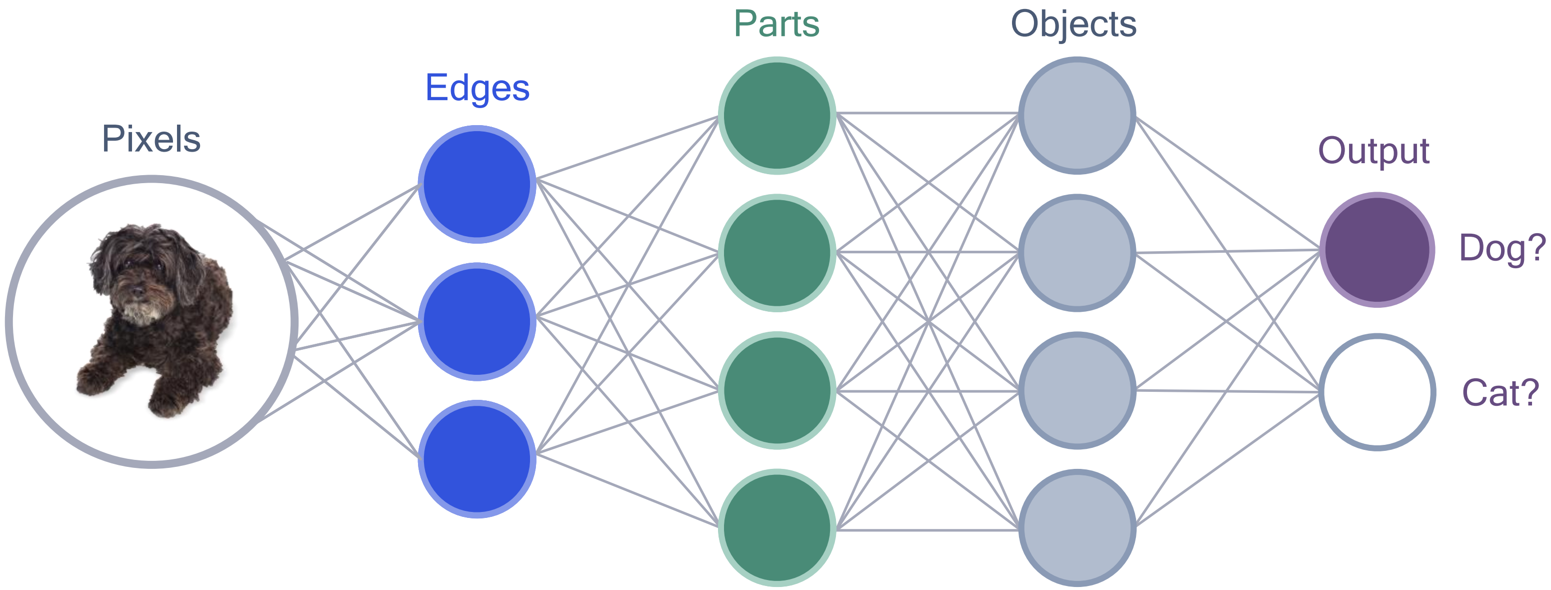
CNNs use too much memory, compute, and energy (today)

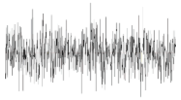
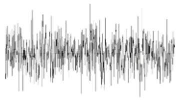
- CNNs do not encode additional symmetries, such as rotation invariance (object may appear in any orientation[1])
- CNNs do not reliably quantify the confidence in a prediction
- CNNs are easy to fool by changing the input only slightly, such as adversarial examples

1. Cohen & Welling 2016

Bayesian deep-learning addresses these challenges

Inspired by brain functionality, introducing noise to neural networks is beneficial

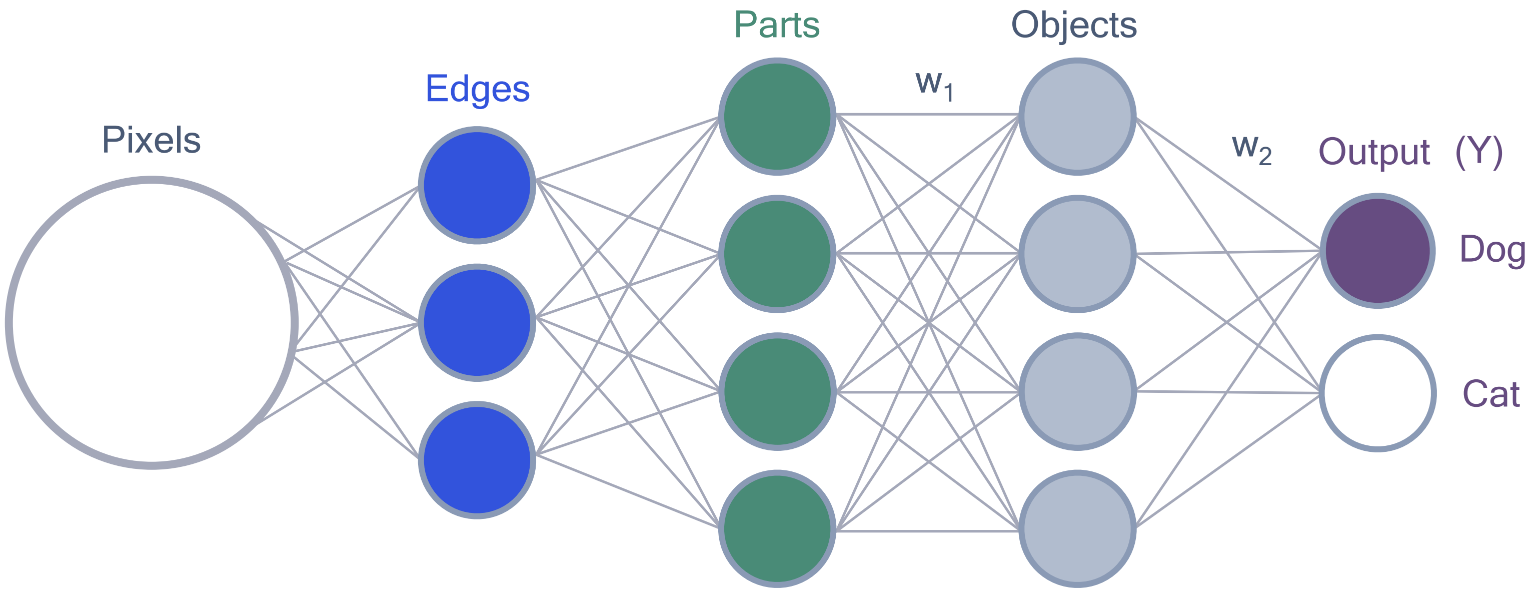


Introduce noise to weights  Noise propagates to activations 

Noise can be a good thing for AI

Compression and quantization

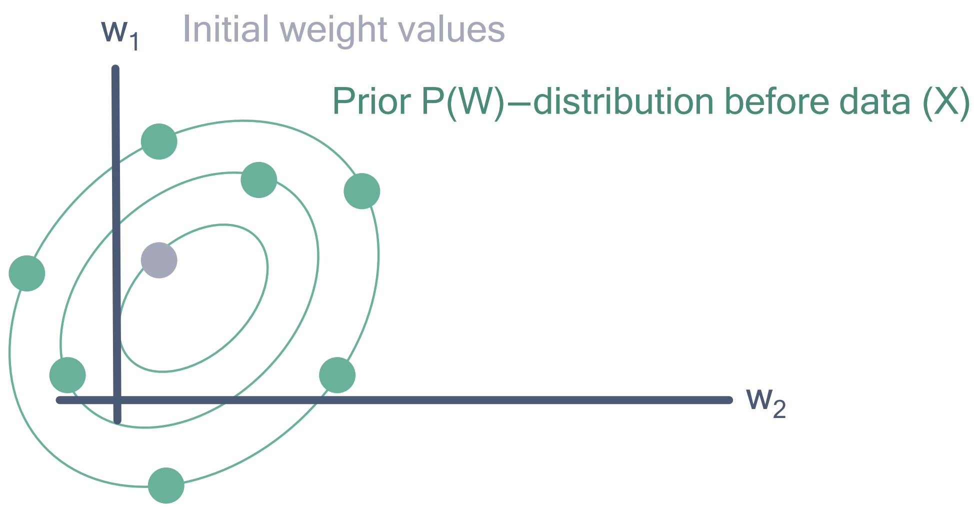
- Reduce complexity of the neural network model
- Reduce bit-width of the parameters and activations
- Save power and improve efficiency



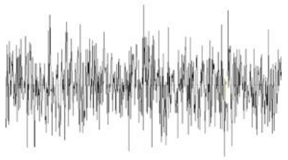
Apply Bayesian deep-learning to shrink the model

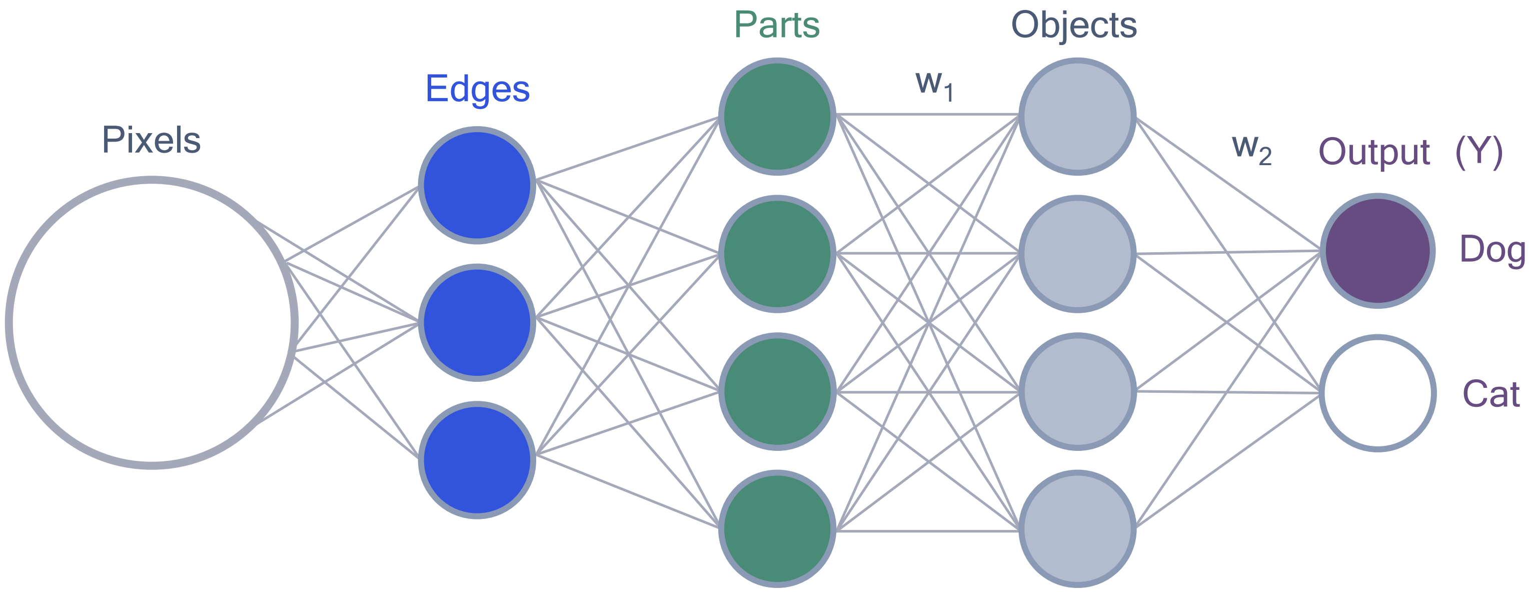
Compression and quantization

- Quantize weights:
  Use lower precision (bit-width)
- Prune activations:
  Reduce number of activation nodes



Introduce noise to parameters

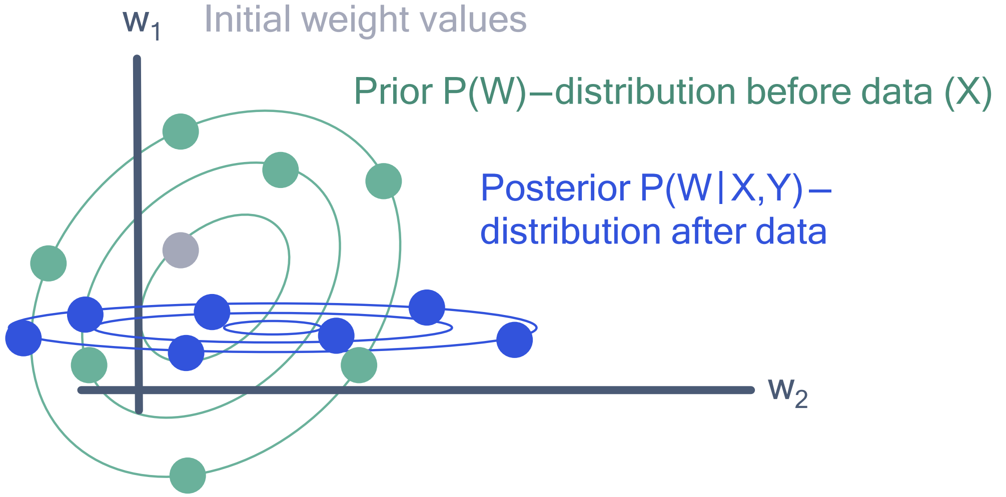




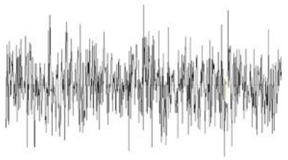
Apply Bayesian deep-learning to shrink the model

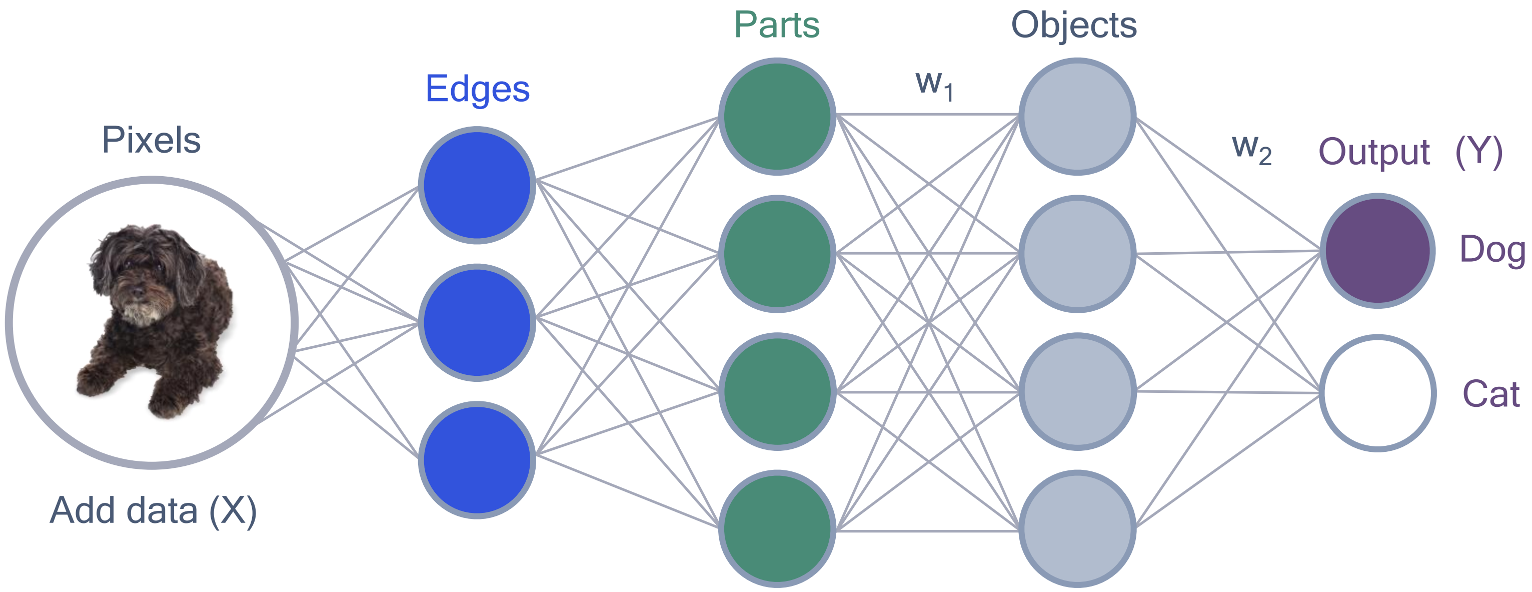
Compression and quantization

- Quantize weights:
  Use lower precision (bit-width)
- Prune activations:
  Reduce number of activation nodes



Introduce noise to parameters

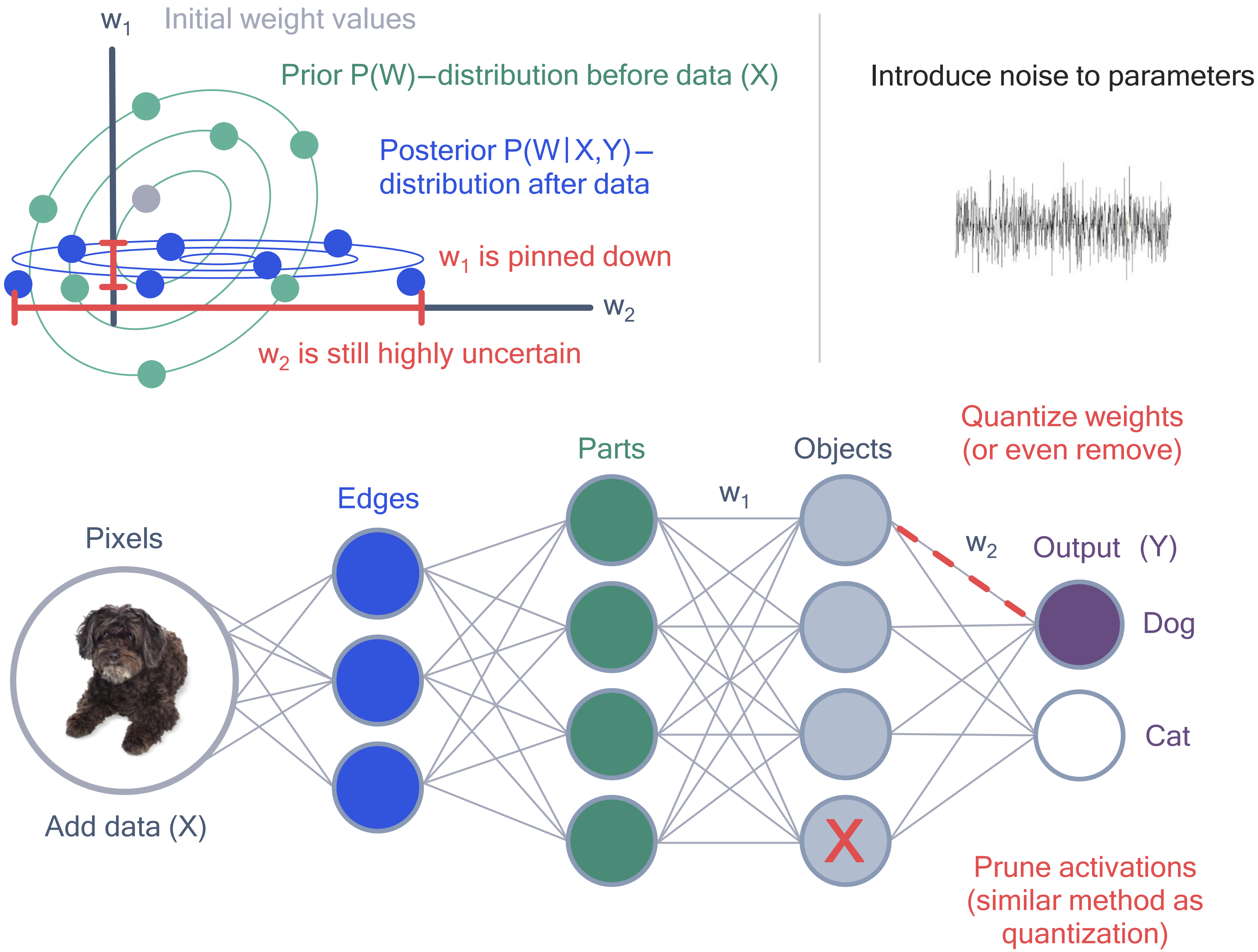




Apply Bayesian deep-learning to shrink the model

Compression and quantization

- Quantize weights:
  Use lower precision (bit-width)
- Prune activations:
  Reduce number of activation nodes



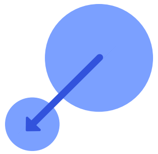
Apply Bayesian deep-learning to shrink the model

Compression and quantization

- Quantize weights:
  Use lower precision (bit-width)
- Prune activations:
  Reduce number of activation nodes

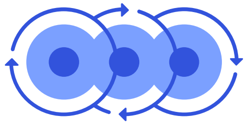
Bayesian deep learning provides broad benefits

A powerful tool to address a variety of deep learning challenges



Compression and quantization

Quantize parameters and activations, prune model components



Regularization and generalization

Avoid overfitting data; choose the simplest model to explain observations (Occam's razor)

Confidence estimation

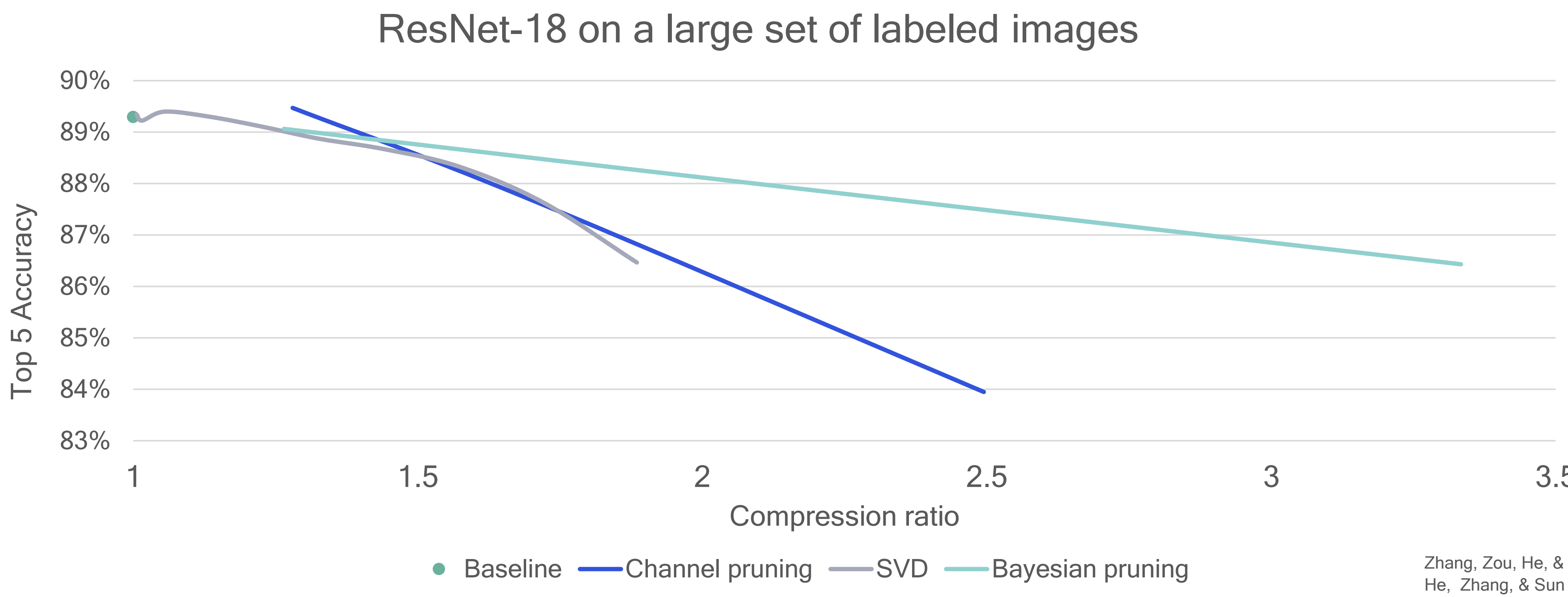
Generate the confidence intervals of the predictions

Privacy/adversarial robustness

Avoid storing personal information in parameters, be less sensitive to adversarial attacks

Applying Bayesian deep learning to real use cases

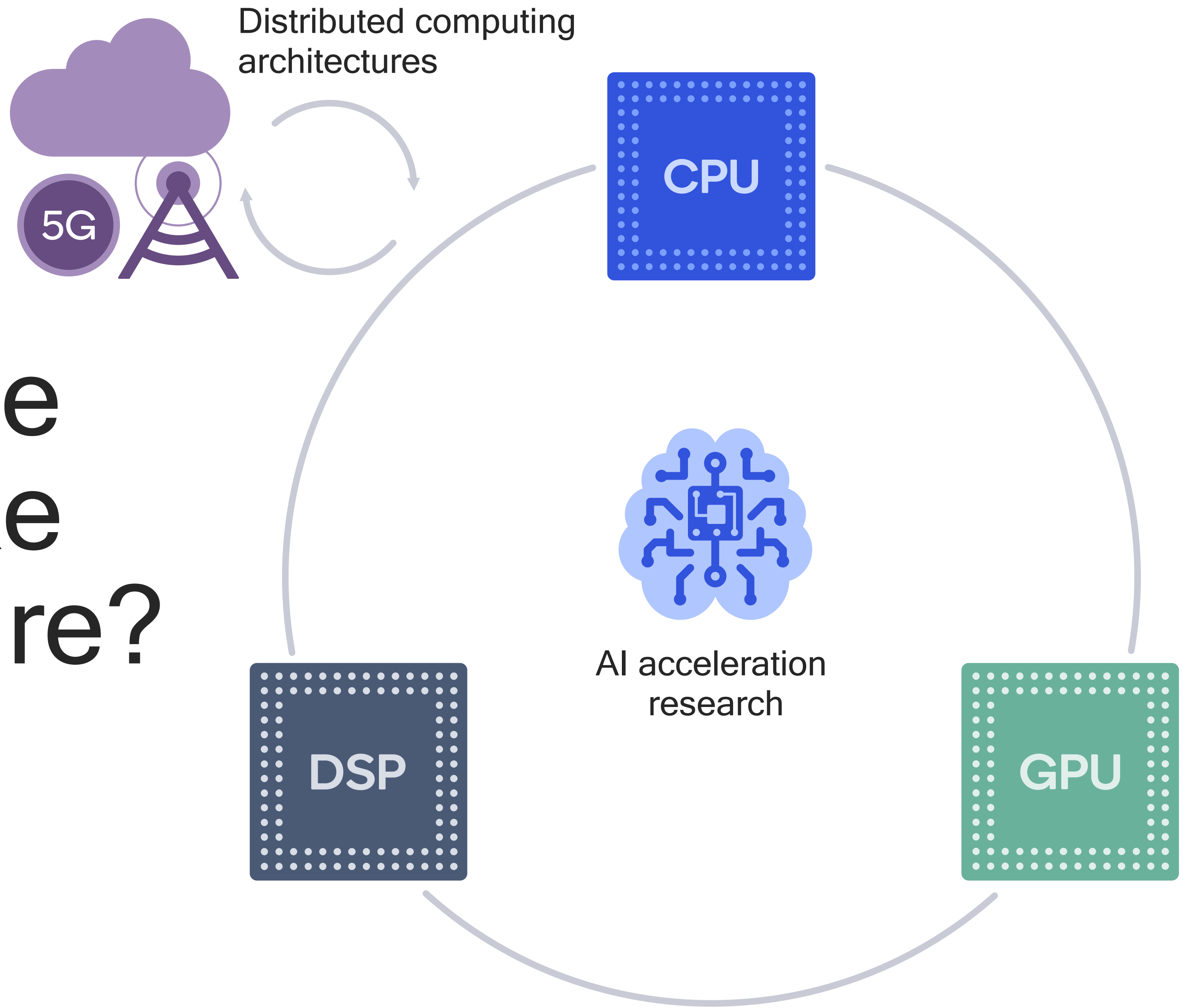
Image classification



Zhang, Zou, He, & Sun 2015 (SVD);
He, Zhang, & Sun 2017 (channel pruning)

3X | compression ratio while maintaining close to the same accuracy

What does the future look like for AI hardware?



AI acceleration research

Focused on fundamentals to accelerate deep learning workloads at low power

Compute architecture

Optimize instruction type, parallelism, and precision of the functional units

Memory hierarchy

Optimize the memory hierarchy to reduce the power consumption of data movement while ensuring performance

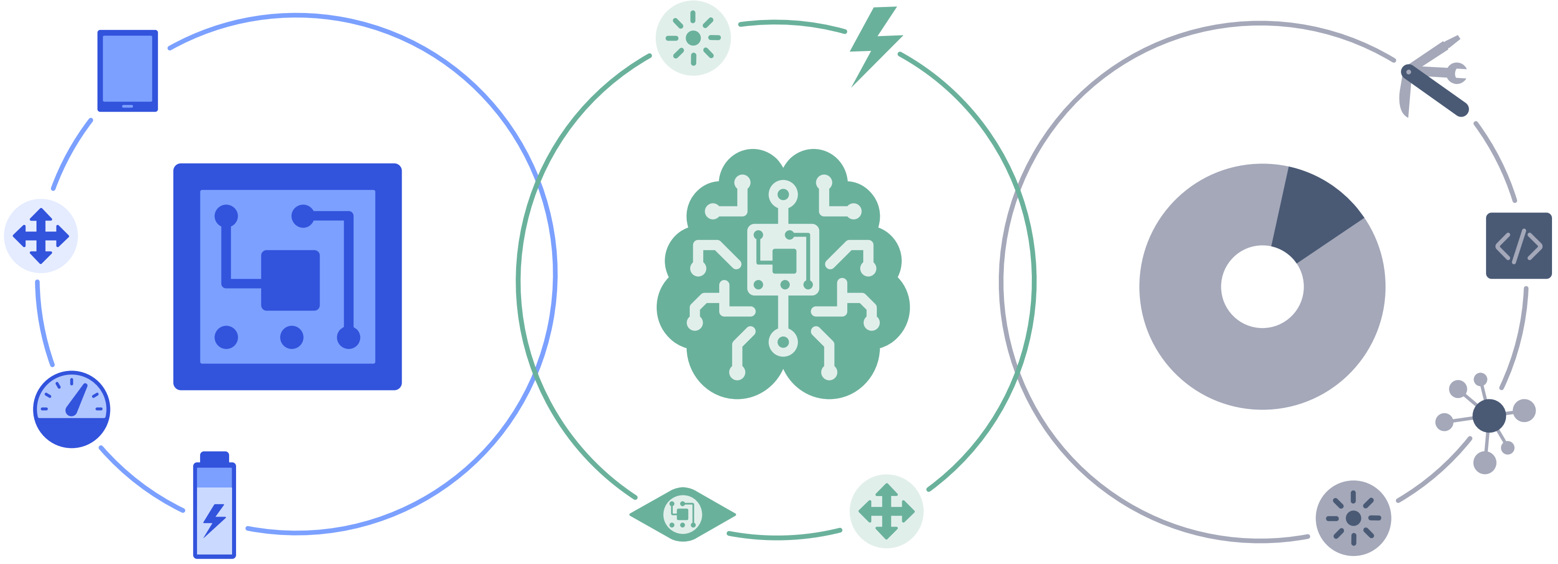
Utilization

Exploit sparsity to improve utilization and reduce power consumption

Balance AI acceleration capabilities across engines (CPU, GPU, DSP)

The approach for making power efficient on-device AI

Focusing on the joint design of algorithms and hardware to achieve high performance



Efficient hardware

Efficient architecture design

Selecting the right compute engine for the right task

Algorithmic advancements

Neural network algorithm design optimized for hardware

Optimization for space and runtime efficiency

Software tools

Software-accelerated run-time for deep learning

Neural processing SDK for model compression and optimization

AI algorithms and hardware need to be energy efficient

We are a leader in Bayesian deep learning and its applications to model compression and quantization

---

We are doing fundamental research on AI algorithms, software, and hardware to maximize power efficiency

Qualcomm

Thank you

Follow us on:

For more information, visit us at:

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