Model Optimizations: Pruning

Model Optimization Use Cases

- **Reducing latency and cost** for inference for both cloud and edge devices (e.g. mobile, IoT)

Model Optimization Use Cases

- **Reducing latency and cost** for inference for both cloud and edge devices (e.g. mobile, IoT)
- **Deploying models on edge devices** with restrictions on processing, memory and/or power-consumption

Model Optimization Use Cases

- **Reducing latency and cost** for inference for both cloud and edge devices (e.g. mobile, IoT)
- **Deploying models on edge devices** with restrictions on processing, memory and/or power-consumption
- **Reducing payload size** for over-the-air model updates

Model Optimization Use Cases

- **Reducing latency and cost** for inference for both cloud and edge devices (e.g. mobile, IoT)
- **Deploying models on edge devices** with restrictions on processing, memory and/or power-consumption
- **Reducing payload size** for over-the-air model updates
- Enabling execution on hardware restricted-to or optimized-for fixed-point operations
- Optimizing models for special purpose hardware accelerators.

The MLOps **Personas**

ML Engineer

ML Researcher

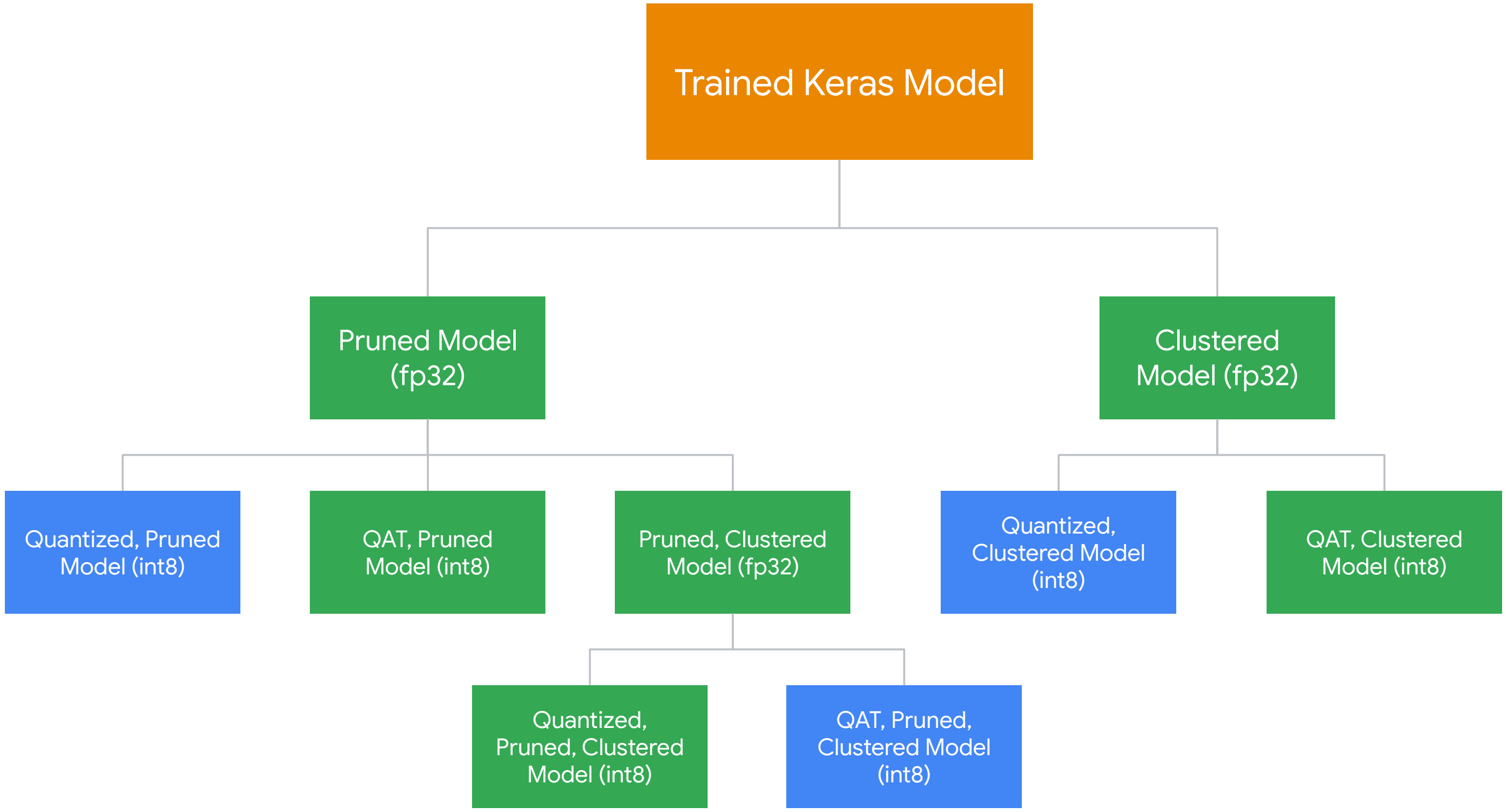
Data Scientist

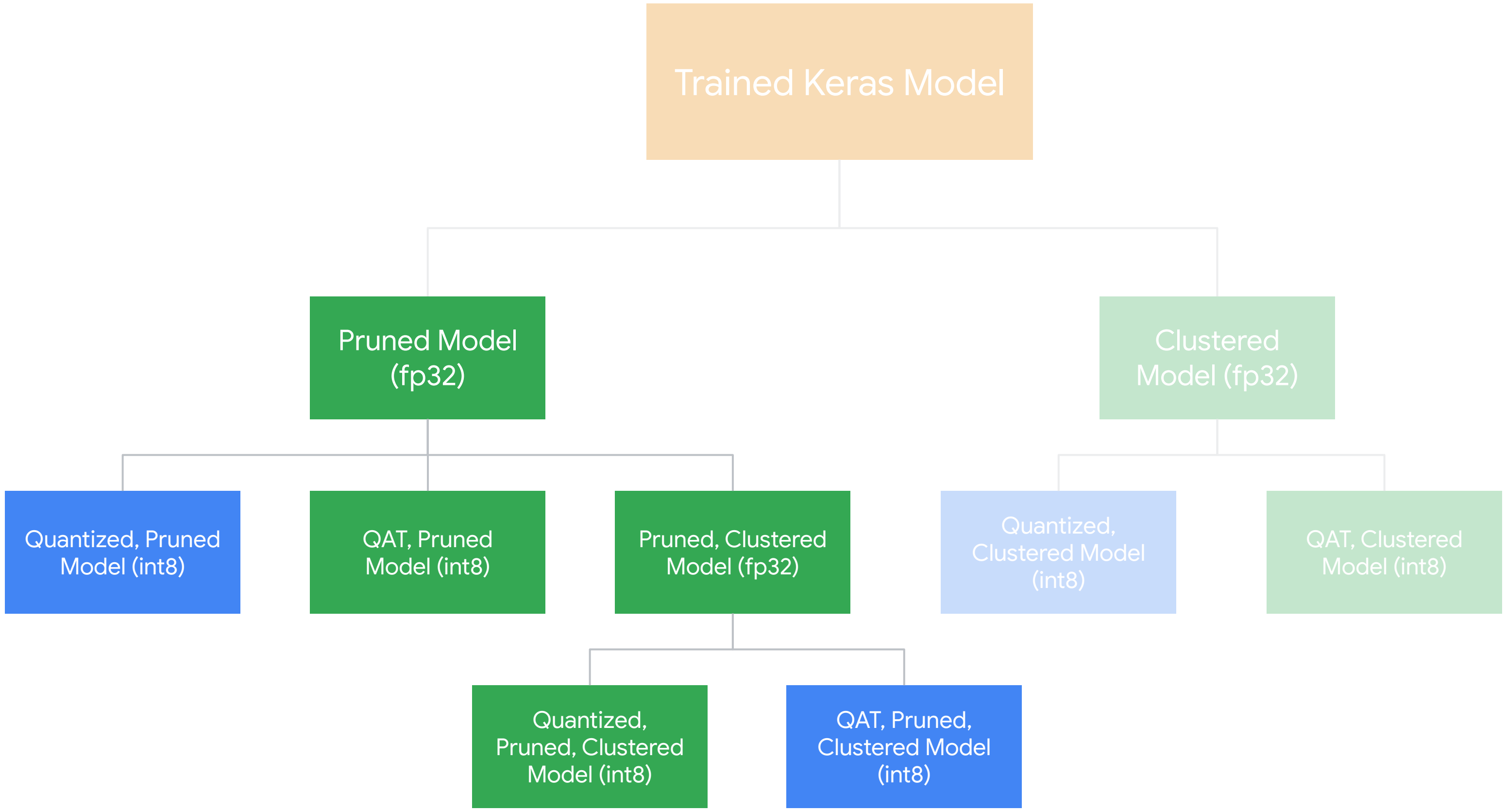
Data Engineer

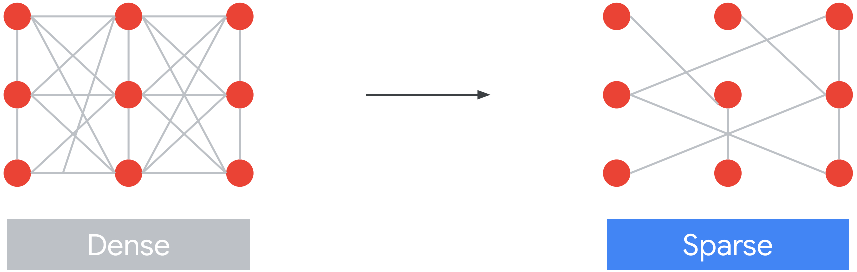
Software Engineer

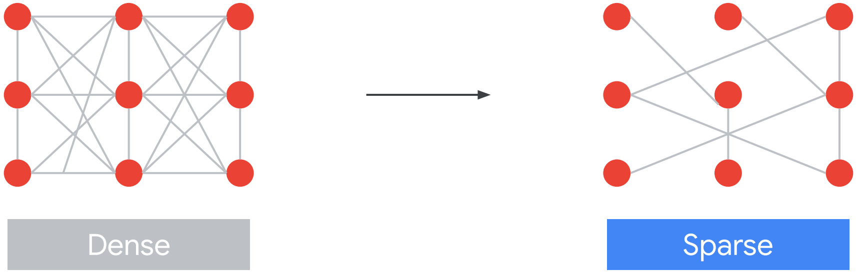
DevOps

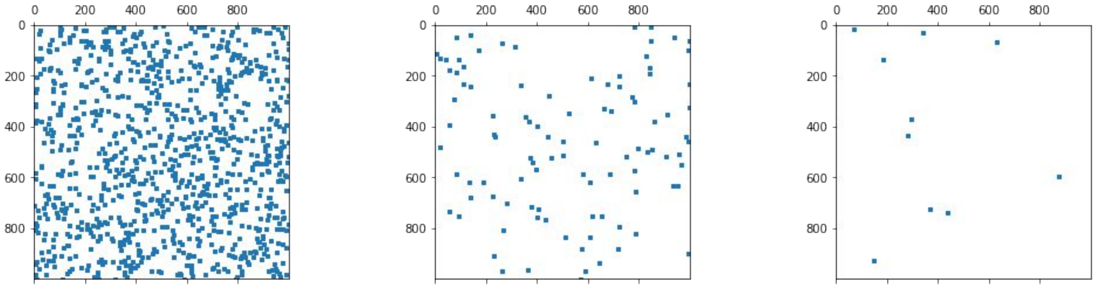
Business Analyst

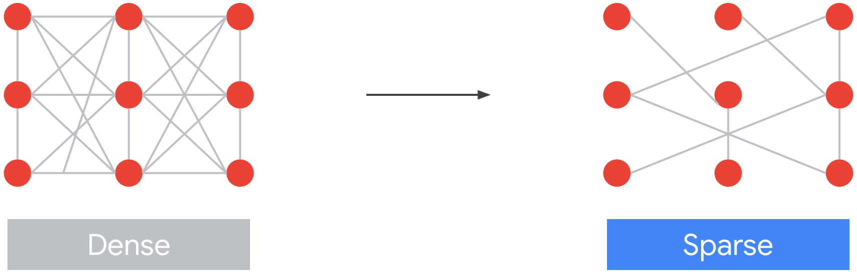

Trained Keras Model
Pruned Model (fp32)
Clustered Model (fp32)
Quantized, Pruned Model (int8)
QAT, Pruned Model (int8)
Pruned, Clustered Model (fp32)
Quantized, Clustered Model (int8)
QAT, Clustered Model (int8)
Quantized, Pruned, Clustered Model (int8)
QAT, Pruned, Clustered Model (int8)

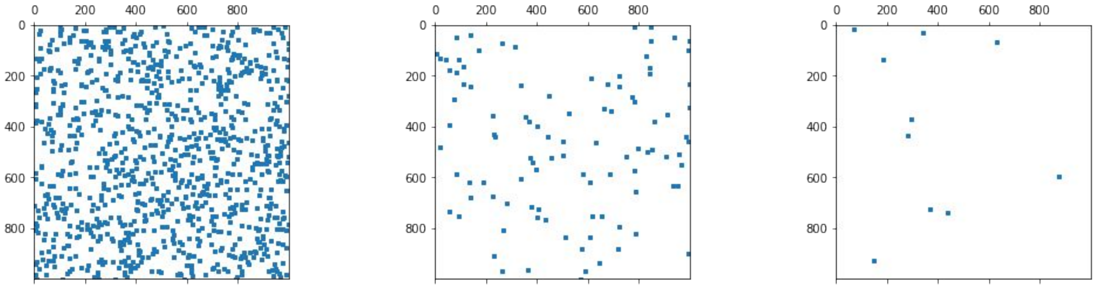

Trained Keras Model
Pruned Model (fp32)
Clustered Model (fp32)
Quantized, Pruned Model (int8)
QAT, Pruned Model (int8)
Pruned, Clustered Model (fp32)
Quantized, Clustered Model (int8)
QAT, Clustered Model (int8)
Quantized, Pruned, Clustered Model (int8)
QAT, Pruned, Clustered Model (int8)


Dense
Sparse


Dense
Sparse


0
200
400
600
800

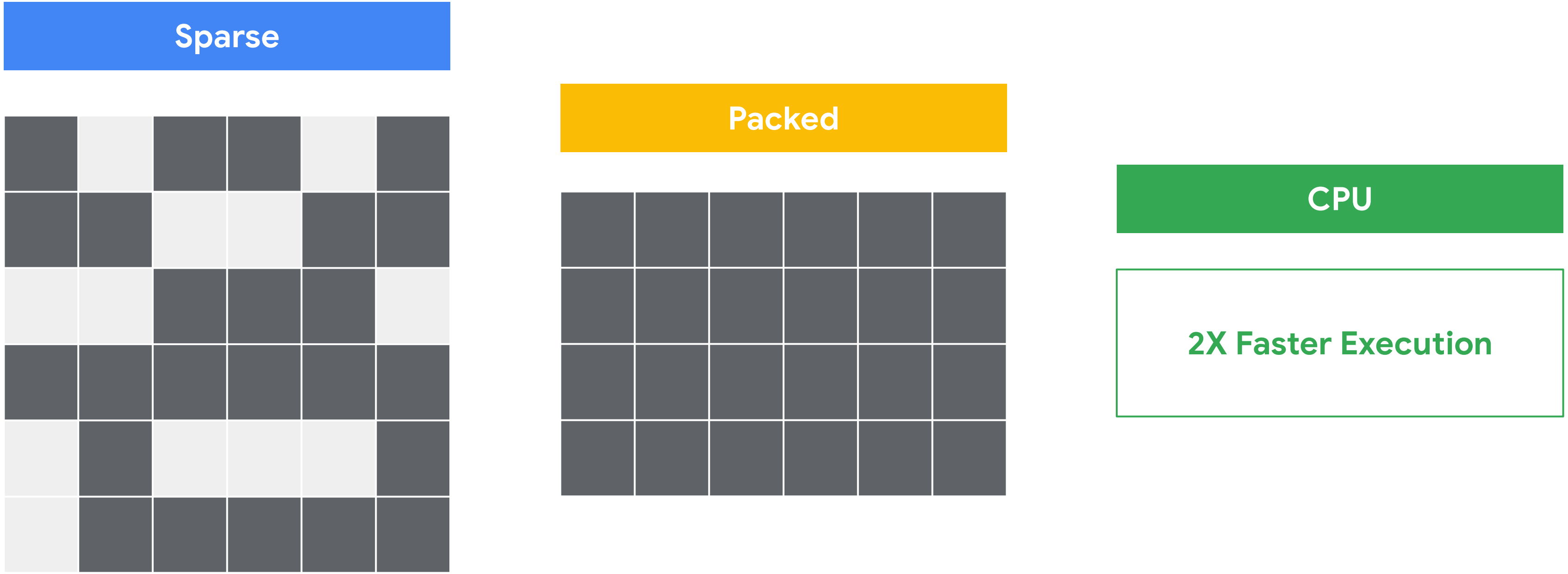




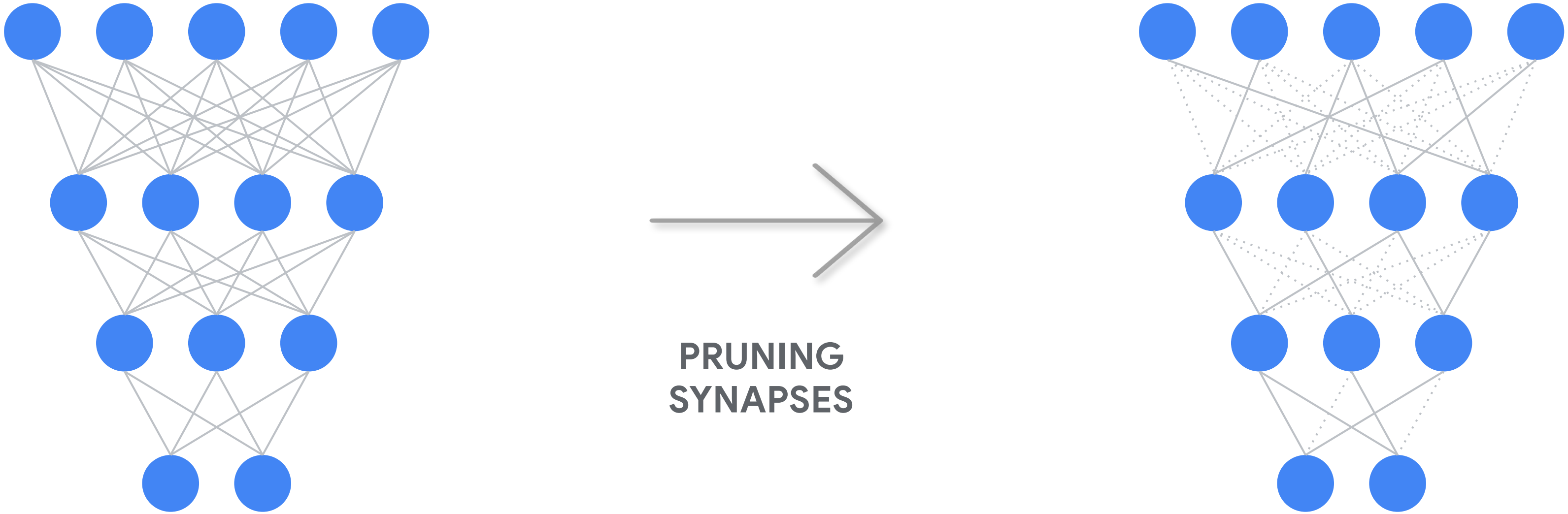
3	4	8
7	6	3
5	5	1

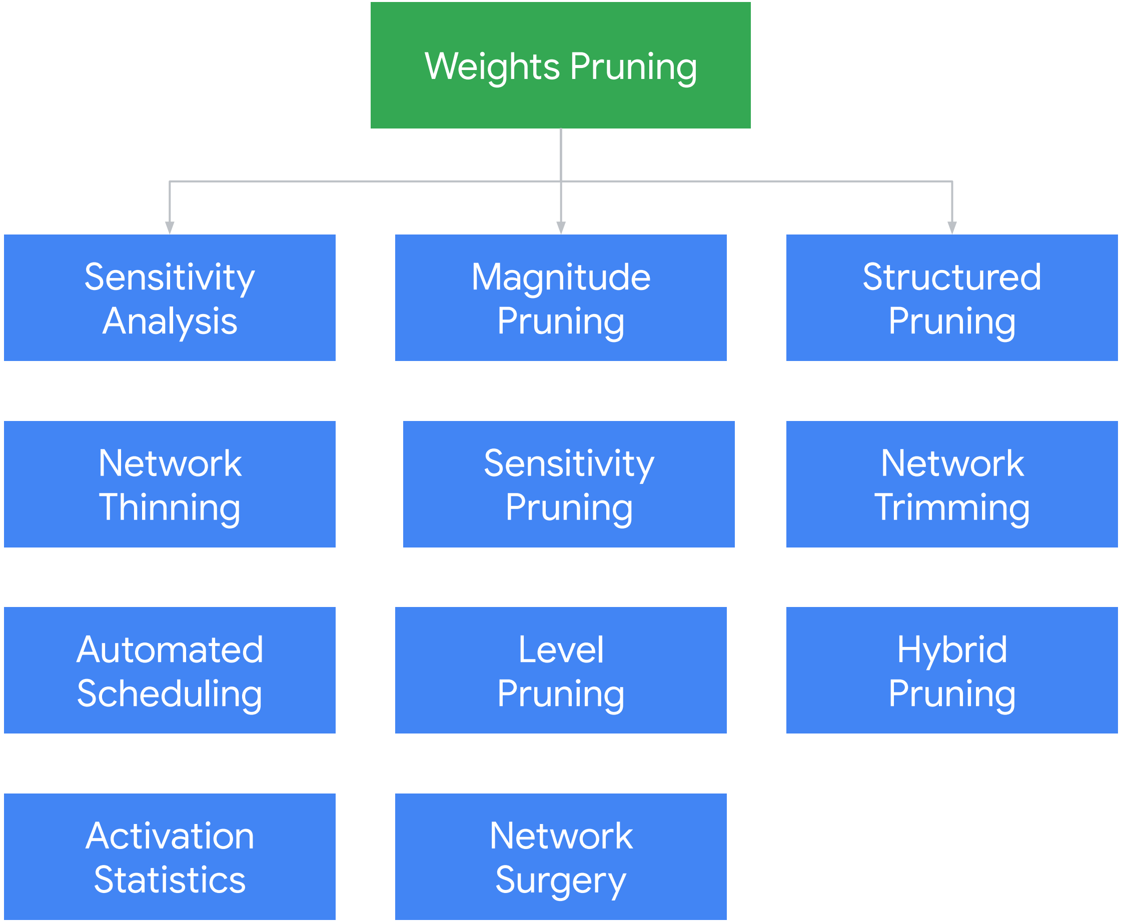
0	4	8
7	6	3
0	5	0

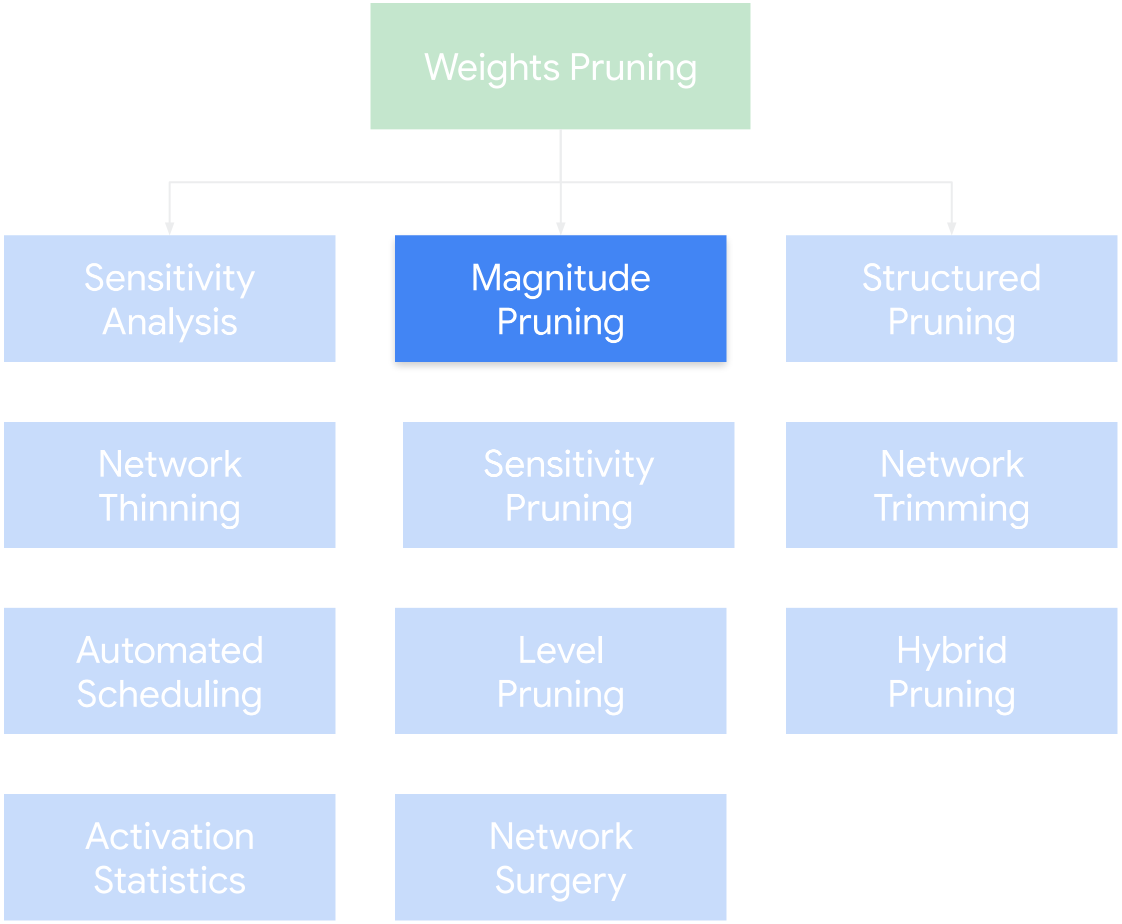
0	4	0
7	0	3
0	5	0


Sparse
Packed
CPU
2X Faster Execution

Pruning




Weights Pruning
Sensitivity Analysis
Magnitude Pruning
Structured Pruning
Network Thinning
Sensitivity Pruning
Network Trimming
Automated Scheduling
Level Pruning
Hybrid Pruning
Activation Statistics
Network Surgery


Weights Pruning
Sensitivity Analysis
Magnitude Pruning
Structured Pruning
Network Thinning
Sensitivity Pruning
Network Trimming
Automated Scheduling
Level Pruning
Hybrid Pruning
Activation Statistics
Network Surgery

Magnitude Pruning

- Sparse models are **easier to compress**

$$thresh(w_i) = \begin{cases} w_i : & if\ |w_i| > \lambda \\ 0 : & if\ |w_i| \leq \lambda \end{cases}$$

Magnitude Pruning

- Sparse models are **easier to compress**

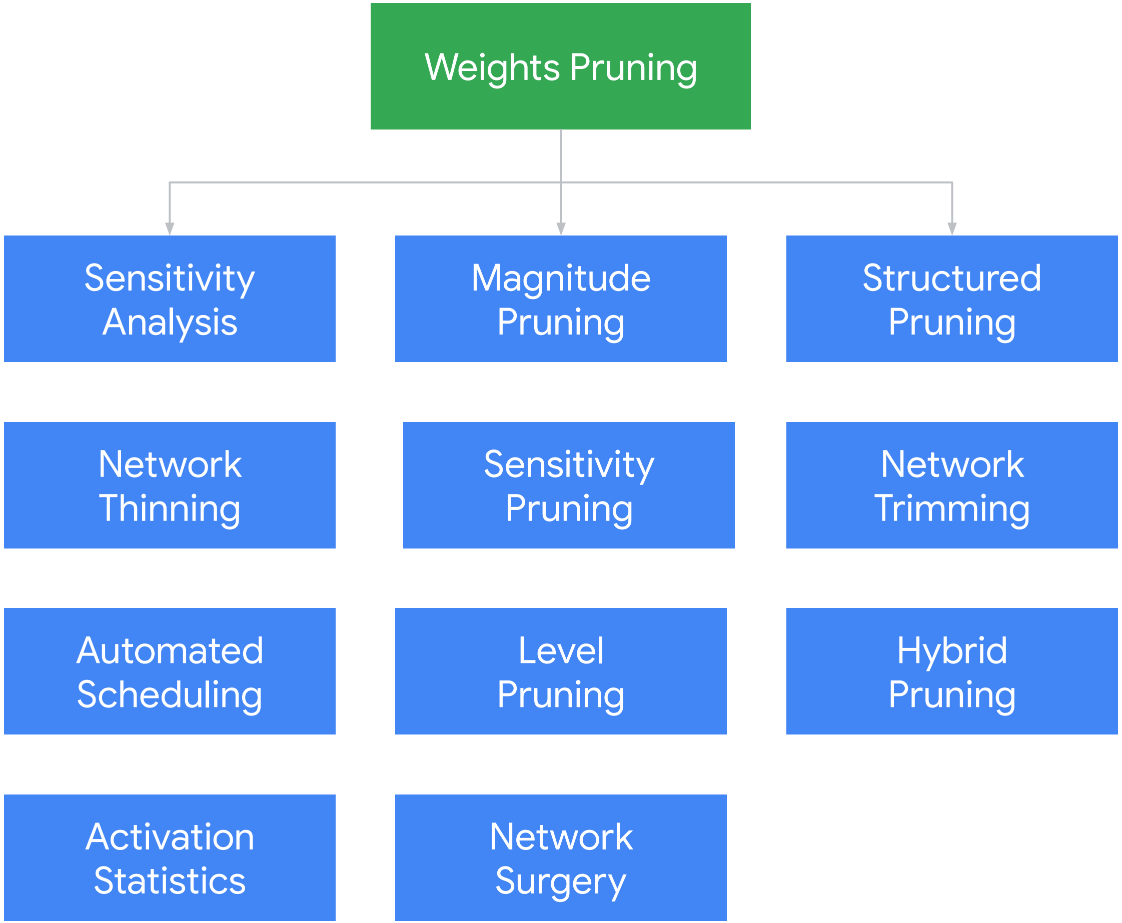
- We can **skip the zeroes during inference for latency** improvements

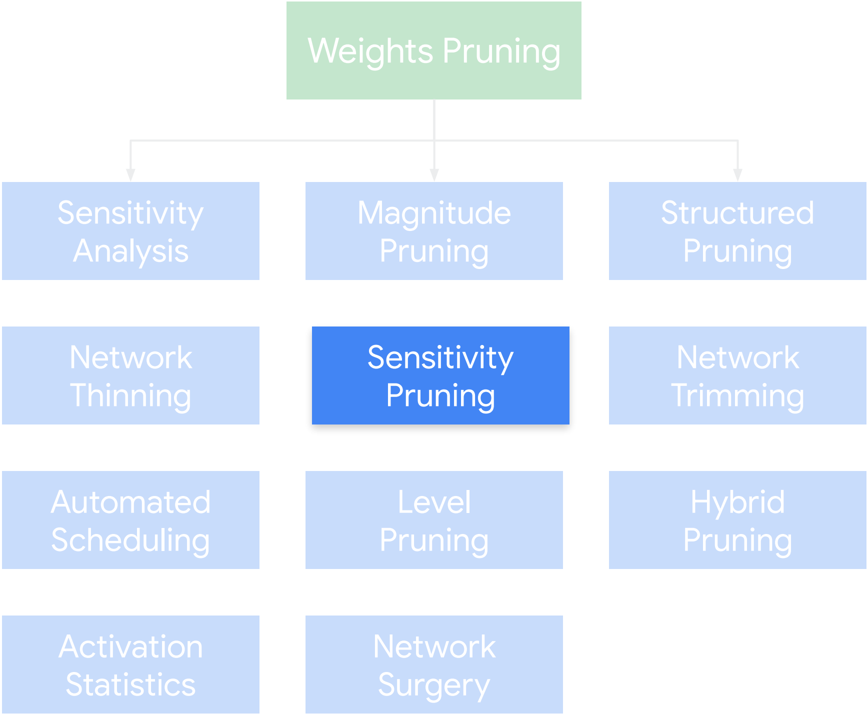
$$thresh(w_i) = \begin{cases} w_i : & if \ |w_i| > \lambda \\ 0 : & if \ |w_i| \leq \lambda \end{cases}$$

Magnitude Pruning

- Sparse models are **easier to compress**
- We can **skip the zeroes during inference for latency** improvements
- Up to **6x improvement**

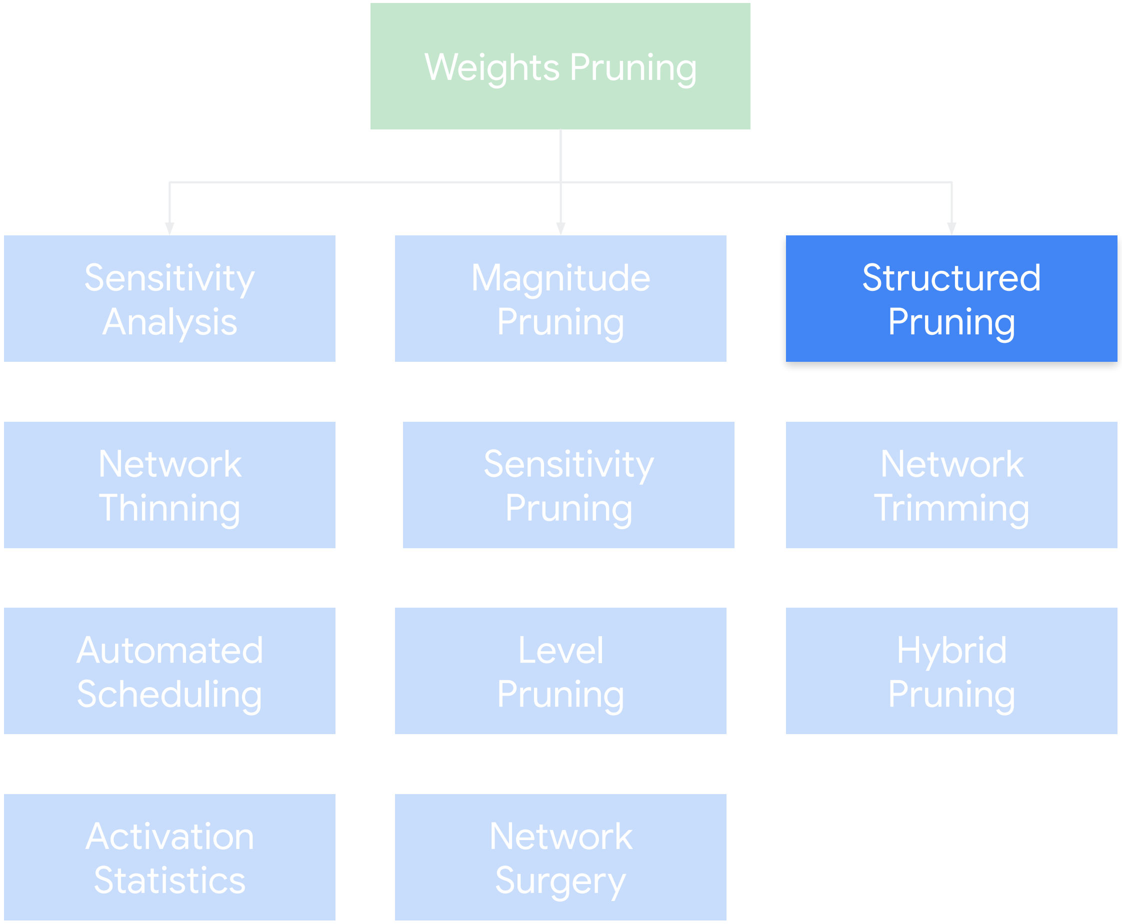
$$thresh(w_i) = \begin{cases} w_i : & if\ |w_i| > \lambda \\ 0 : & if\ |w_i| \leq \lambda \end{cases}$$


Weights Pruning
Sensitivity Analysis
Magnitude Pruning
Structured Pruning
Network Thinning
Sensitivity Pruning
Network Trimming
Automated Scheduling
Level Pruning
Hybrid Pruning
Activation Statistics
Network Surgery

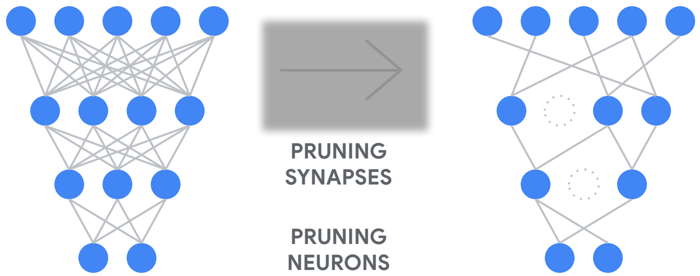


$$thresh(w_i) = \begin{cases} w_i : & if \; |w_i| > \lambda \\ 0 : & if \; |w_i| \leq \lambda \end{cases}$$

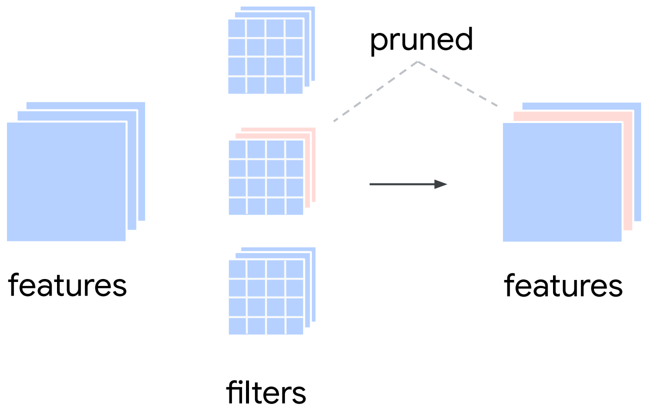
$\lambda = s * \sigma_l$ *where* σ_l *is the std of layer* l *as measured on the dense model*


Weights Pruning
Sensitivity Analysis
Magnitude Pruning
Structured Pruning
Network Thinning
Sensitivity Pruning
Network Trimming
Automated Scheduling
Level Pruning
Hybrid Pruning
Activation Statistics
Network Surgery

Unstructured Pruning



Structured Pruning



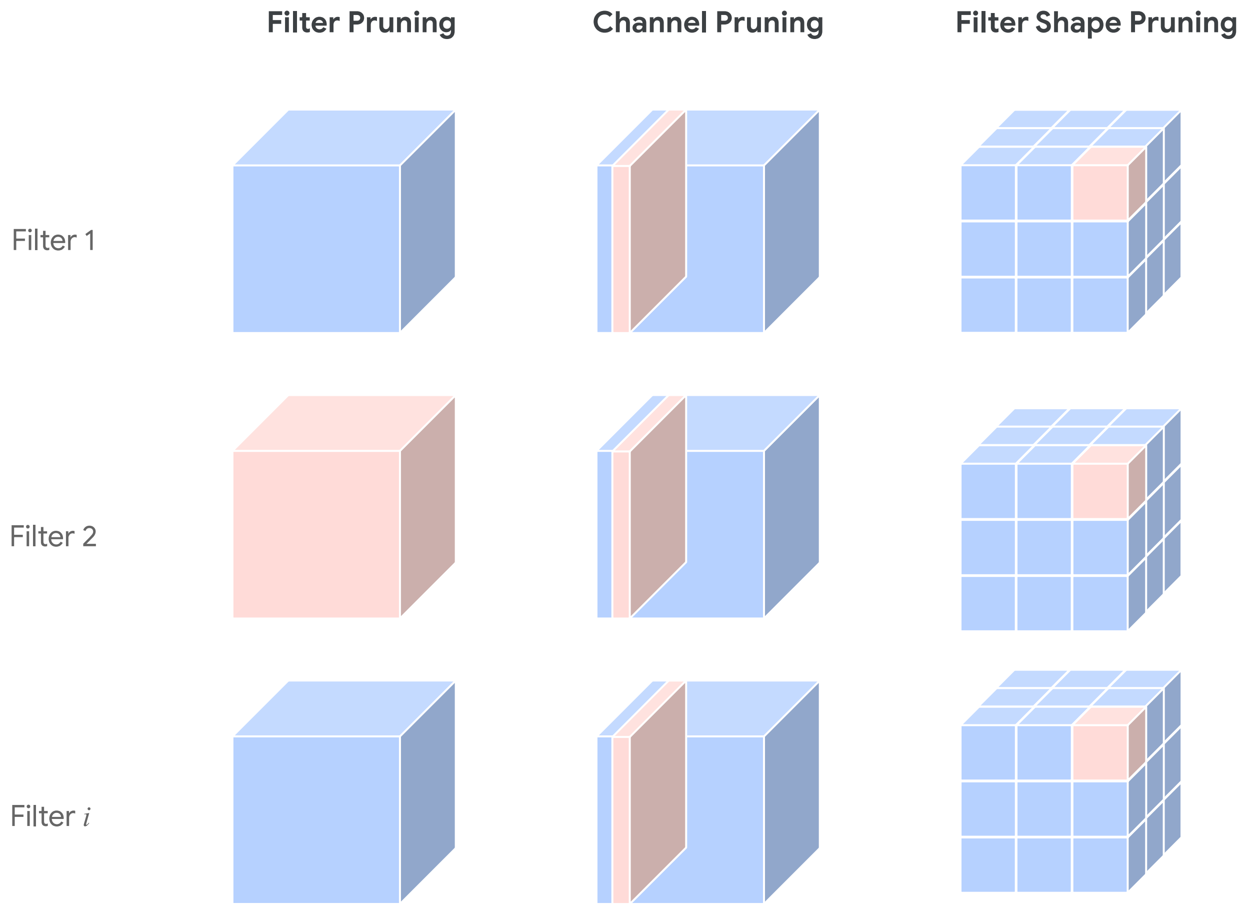

Filter Pruning
Channel Pruning
Filter Shape Pruning
Filter 1
Filter 2
Filter i

Image Classification

Model	Non-sparse Top-1 Accuracy	Sparse Accuracy	Sparsity
InceptionV3	78.1%	78.0%	50%
		76.1%	75%
		74.6%	87.5%
MobilenetV1 224	71.04%	70.84%	50%

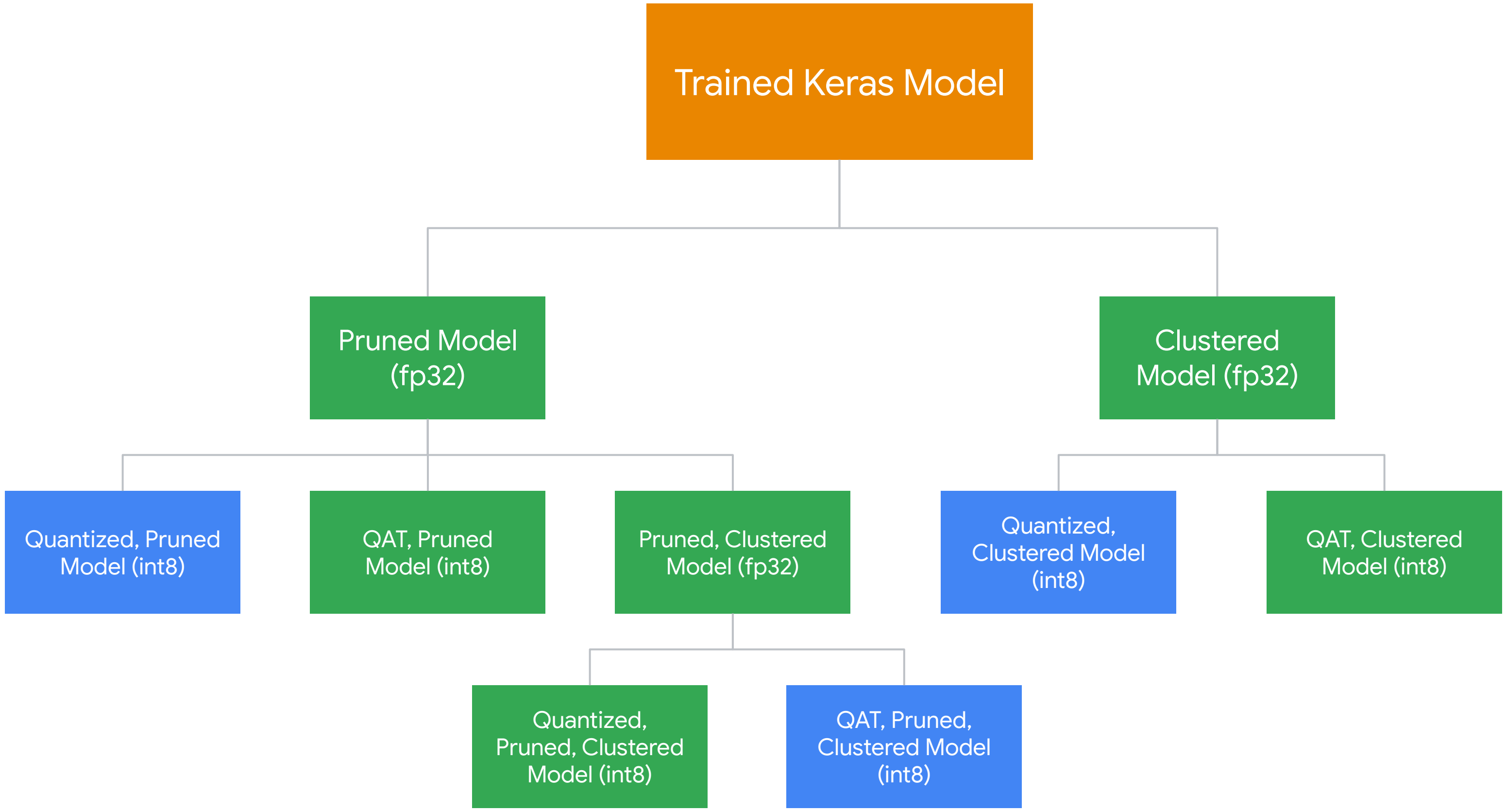
The models were tested on Imagenet.

Language Translation

Model	Non-sparse BLEU	Sparse BLEU	Sparsity
GNMT EN-DE	26.77	26.86	80%
		26.52	85%
		26.19	90%
GNMT DE-EN	29.47	29.50	80%
		29.24	85%
		28.81	90%

Keyword Spotting

Model	Non-sparse Accuracy	Structured Sparse Accuracy (2 by 4 pattern)	Random Sparse Accuracy (target sparsity 50%)
DS-CNN-L	95.23	94.33	94.84


Trained Keras Model
Pruned Model (fp32)
Clustered Model (fp32)
Quantized, Pruned Model (int8)
QAT, Pruned Model (int8)
Pruned, Clustered Model (fp32)
Quantized, Clustered Model (int8)
QAT, Clustered Model (int8)
Quantized, Pruned, Clustered Model (int8)
QAT, Pruned, Clustered Model (int8)