

# tinyML<sup>®</sup> EMEA

*Enabling Ultra-low Power Machine Learning at the Edge*

June 26 - 28, 2023



[www.tinyML.org](http://www.tinyML.org)

# Are Hyperparameters Overrated?

## From large models to tinyML



**Presenter: Jonna Matthiesen**

Master Thesis Student / Deep Learning Researcher  
Embedl

**Andreas Ask**

Deep Learning Researcher

**Daniel Ödman**

Deep Learning Researcher

# The Promise of Deep Learning

Deep Learning holds great promise but ...

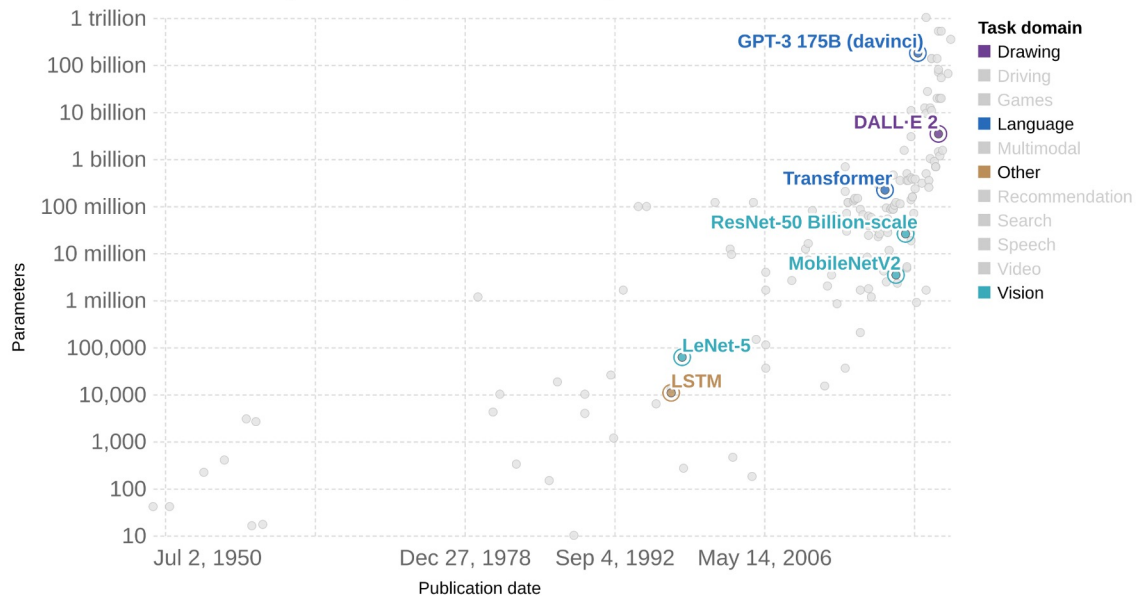
Efficiency is key!

# Bigger and Bigger!

## Number of parameters in notable artificial intelligence systems

Our World  
in Data

Parameters are variables in an AI system whose values are adjusted during training to establish how input data gets transformed into the desired output; for example, the connection weights in an artificial neural network.



Source: Sevilla et al. (2023)

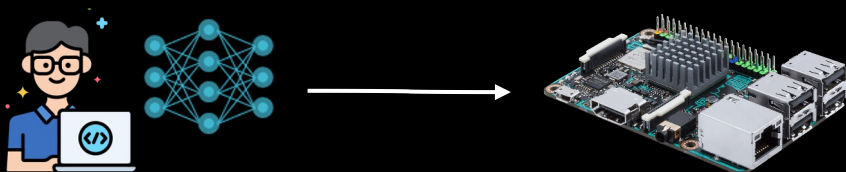
OurWorldInData.org/artificial-intelligence • CC BY

Note: Parameters are estimated based on published results in the AI literature and come with some uncertainty. The authors expect the estimates to be correct within a factor of 10.

# Bigger is better?

With the use of powerful GPUs we can train bigger models but ...

*What about deployment?*



To deploy on the edge we need:

- **Smaller** models
- ... and **faster** models
- ... that are still **accurate**

# About Embedl

---

- Specialised in *optimization of Deep Learning models*
- *Source code* available - Not a black box!
- *Empower* teams of data scientists and DL engineers with powerful optimization *algorithms and tools*
- Based in the automotive capital of Sweden - Gothenburg

Trusted by:

V O L V O



veoneer

zenseact

tobii

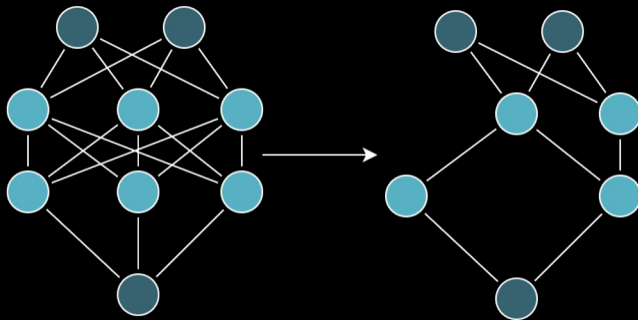


SIEMENS

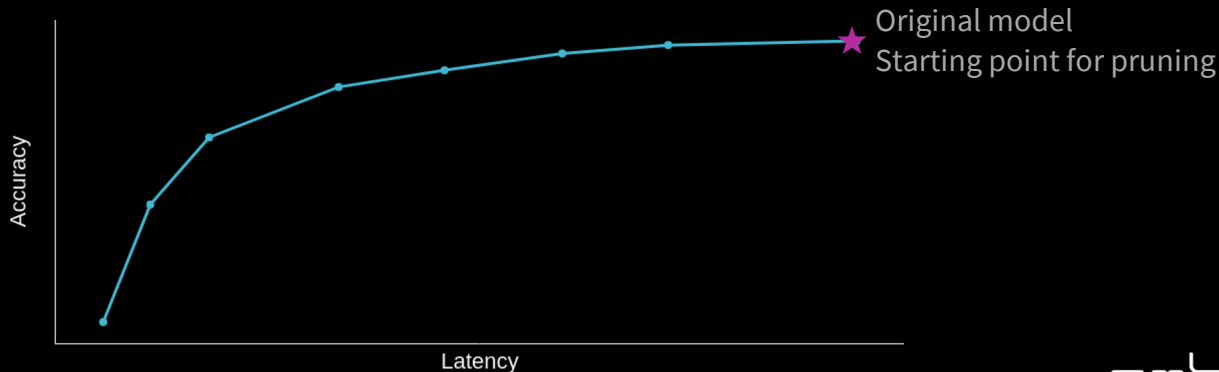
# Focus Today: Deep Neural Network Pruning

Making a model faster with minimal accuracy loss

Removing redundant and unnecessary connections or structures from the network



- ✓ Latency
- ✓ Power
- ✓ Runtime memory



# Introduction to Pruning

When, how, what, and where to prune?

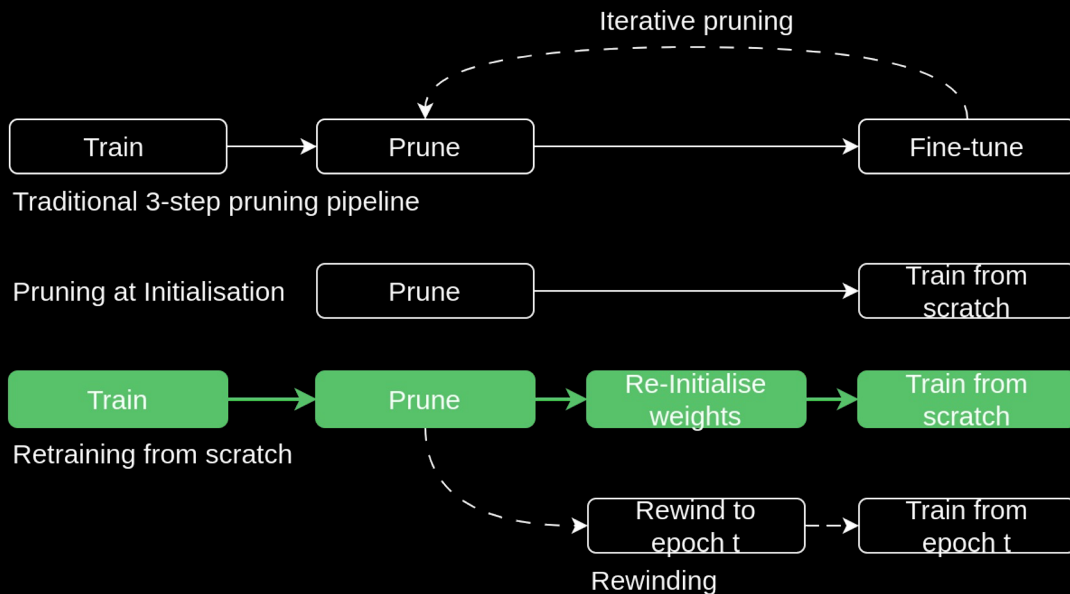


# Introduction to Pruning

## Pipeline

*When to prune?*

Before, during, or after training



# Introduction to Pruning

Pipeline

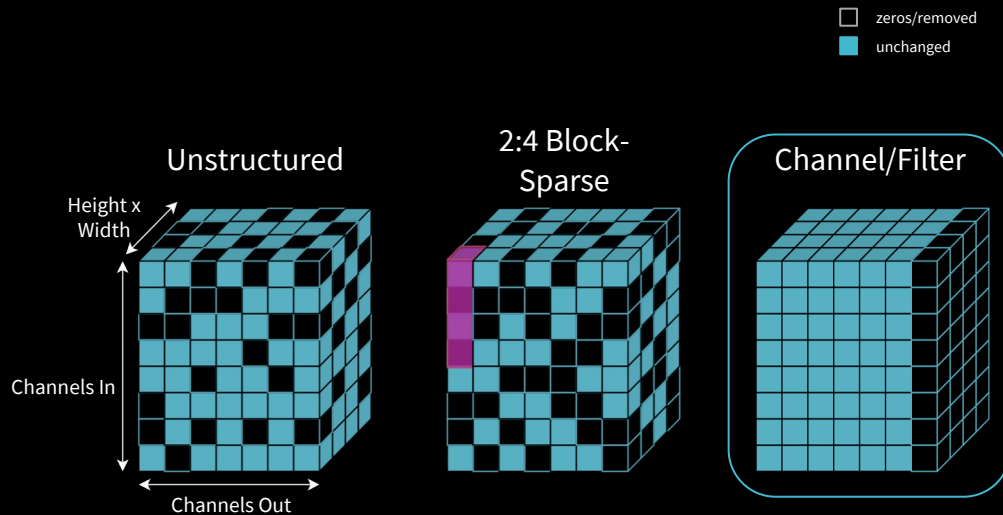
*When to prune?*

Retraining from scratch

Structure

*How to prune?*

Structured vs unstructured



- Structured pruning results in speedup on most hardware
- Today's results focus on **structured pruning**

# Introduction to Pruning

Pipeline

*When to prune?*

Retraining from scratch

Structure

*How to prune?*

Structured pruning

Scoring

*What to prune?*

Parameters scored based on:  
magnitude, importance, gradients etc.

L1-Norm magnitude pruning

- Li, et al., arXiv:1608.08710 (2016) -

Prune filters with the lowest L1-norm

$$\|F_{i,j}\|_1$$

**Assumption:** Weights of small magnitude have a small impact on the performance of a model.

- LeCun, et al., "Optimal brain damage." (1989) -

# Introduction to Pruning

Pipeline

*When to prune?*

Retraining from scratch

Structure

*How to prune?*

Structured pruning

Scoring

*What to prune?*

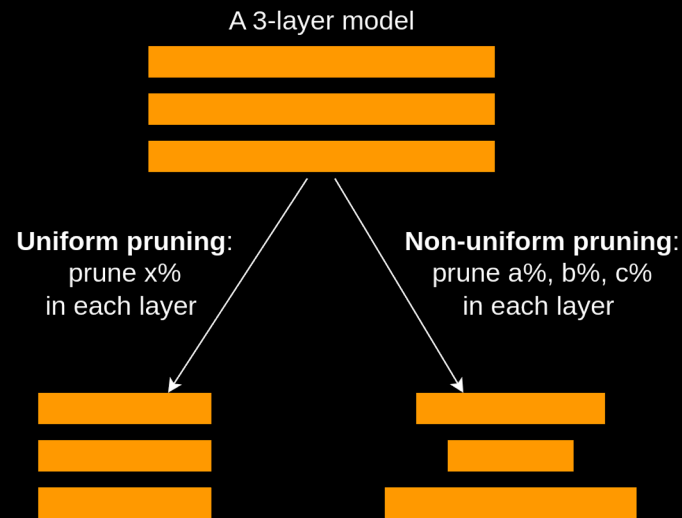
L1-Norm Magnitude Pruning

Method

*Where to prune?*

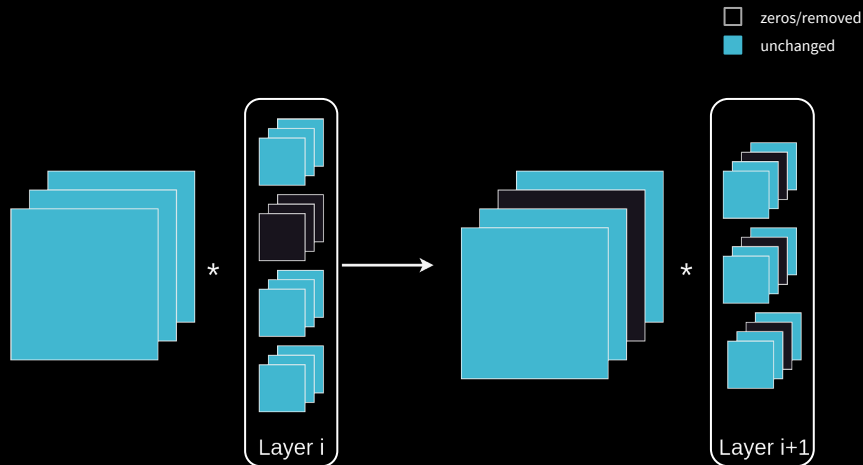
Uniform vs non-uniform pruning

$$\text{Pruning ratio} = \frac{\text{FLOPs of pruned model}}{\text{FLOPs of original model}}$$



# Hyperparameters under Structured Pruning

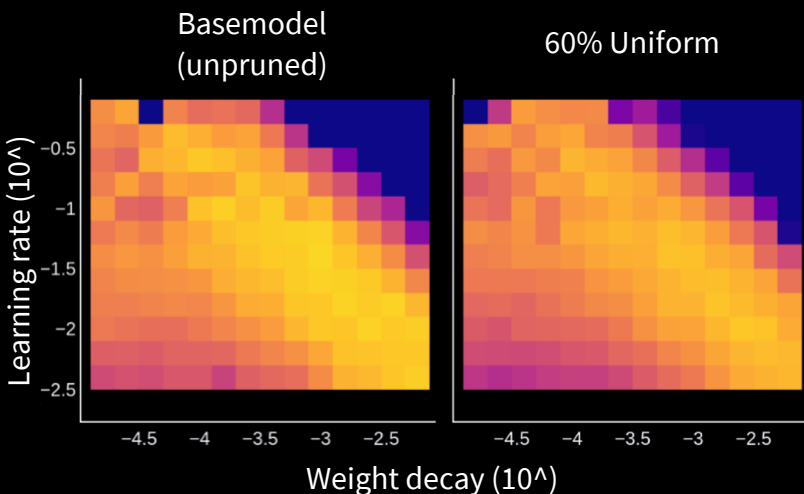
- "A priori, models of different sizes don't have any reason to share the optimal HPs." -  
G. Yang, et al., arXiv:2203.03466 (2022)



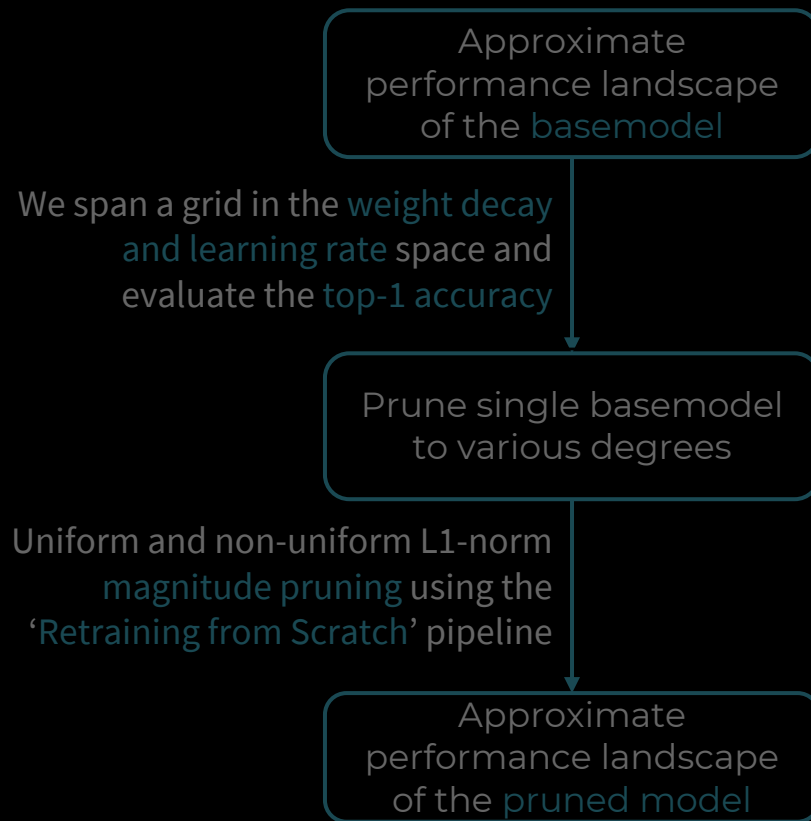
Filter/Channel Pruning changes the architecture of the CNN, making it smaller

What are the effects of structured pruning on the optimal hyperparameters?

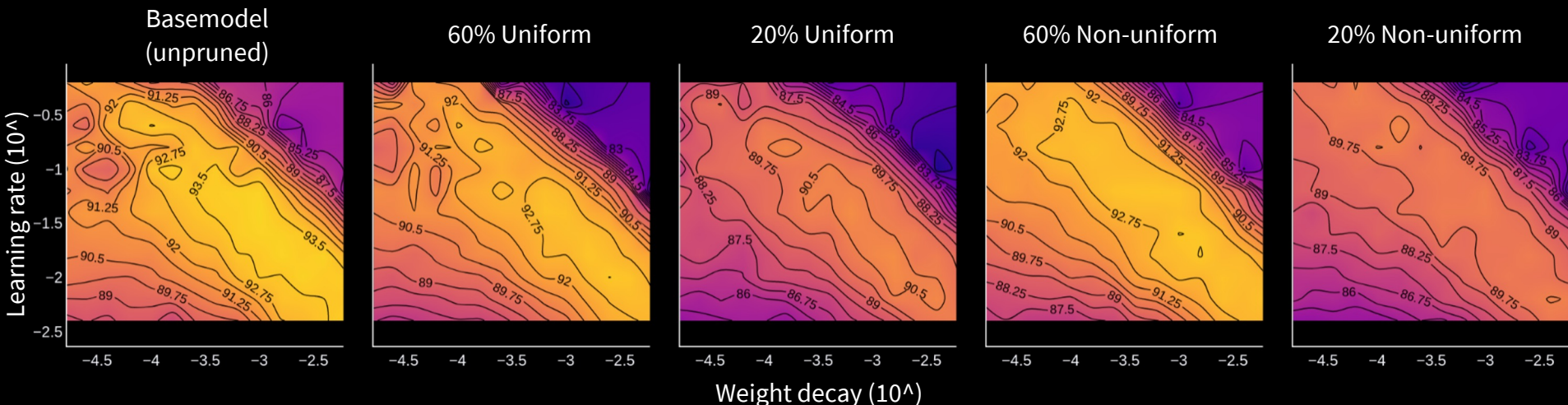
# Experimental Pipeline



- Grid points are interpolated
- Points outside the top 15% (top-1 accuracy) are removed and inter-/extrapolated  
→ smoother visualisation

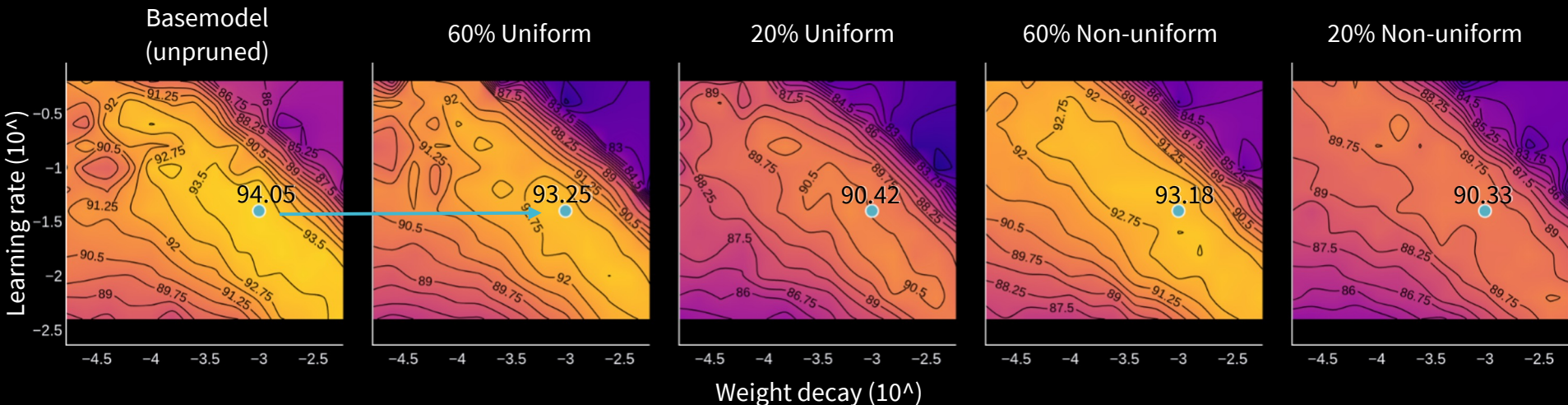


# ResNet-56 on CIFAR-10



- The performance landscape remains relatively stable across different structured pruning methods



# ResNet-56 on CIFAR-10

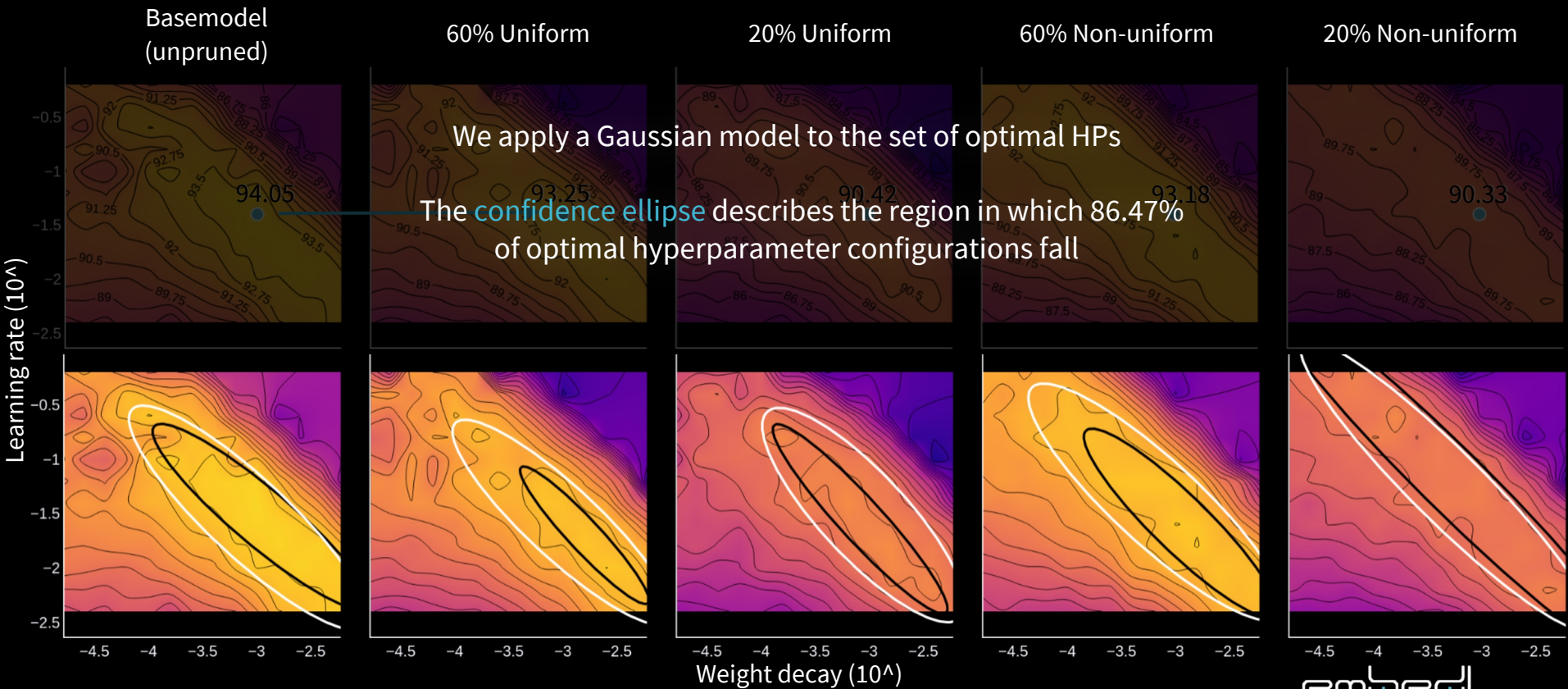


- The performance landscape remains relatively stable across different structured pruning methods
- Optimal hyperparameters of the basemodel can serve as a reasonable starting point for the pruned model





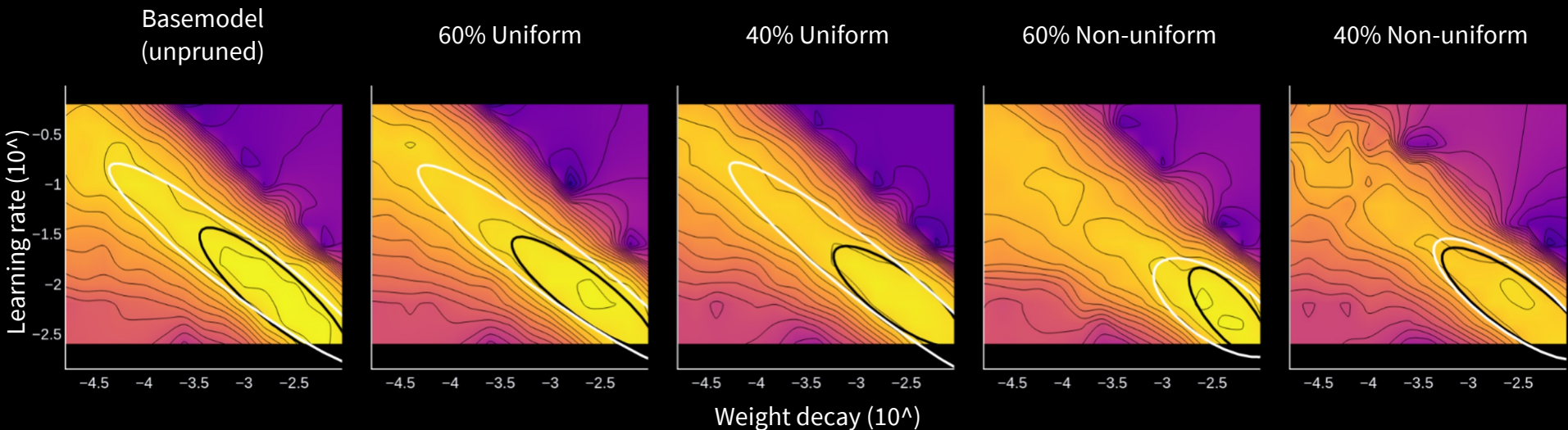
# ResNet-56 on CIFAR-10

-  model performance is within the top 0.5%
-  model performance is within the top 1.0%





# MobileNetV2 on CIFAR-10

-  model performance is within the top 0.5%
-  model performance is within the top 1.0%



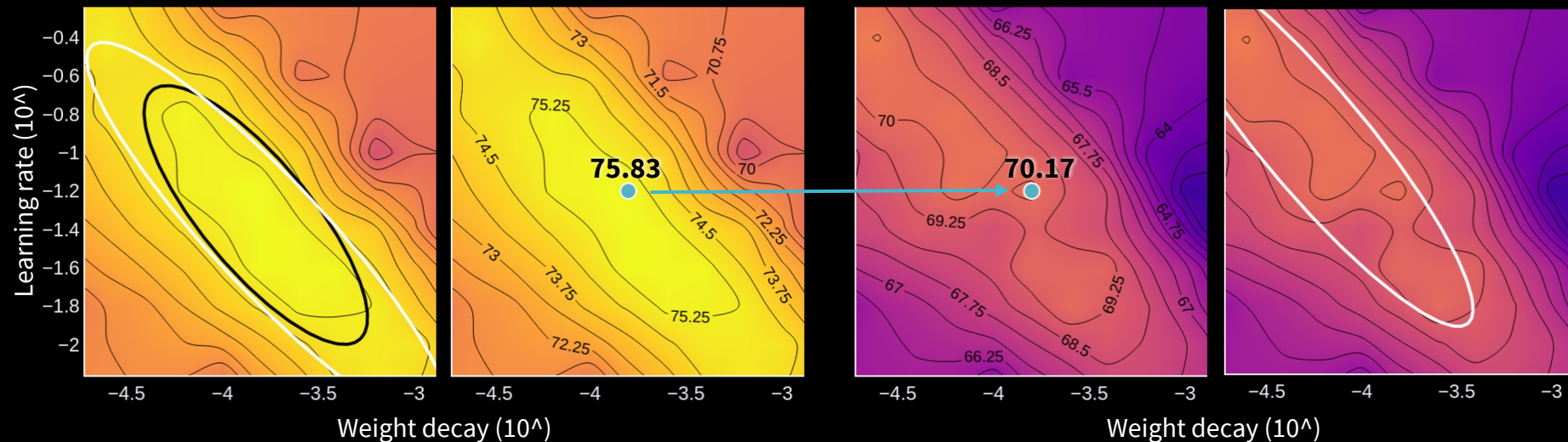
The **confidence ellipse** describes the region in which 86.47% of optimal hyperparameter configurations fall

# ResNet-50 on ImageNet

-  model performance is within the top 0.5%
-  model performance is within the top 1.0%

Basemodel (unpruned)

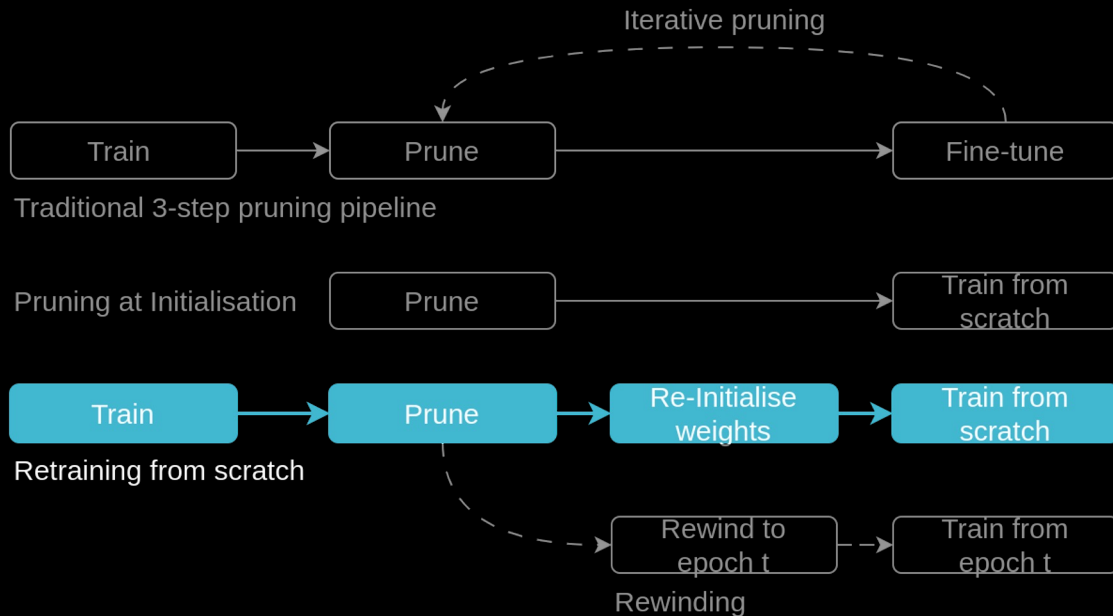
40% Non-uniform



The **confidence ellipse** describes the region in which 86.47% of optimal hyperparameter configurations fall

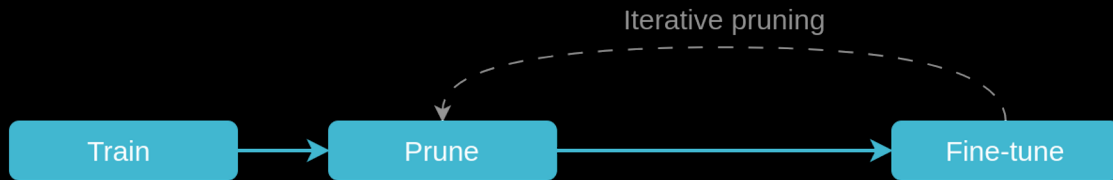
# But wait ...

What about Fine-tuning?



# But wait ...

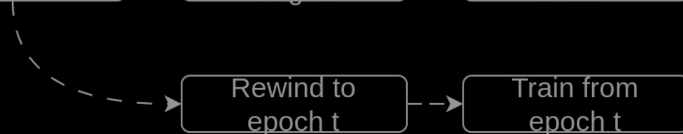
What about Fine-tuning?



Traditional 3-step pruning pipeline



Retraining from scratch



Rewinding

# ResNet-56 on CIFAR-10

60% Non-uniform

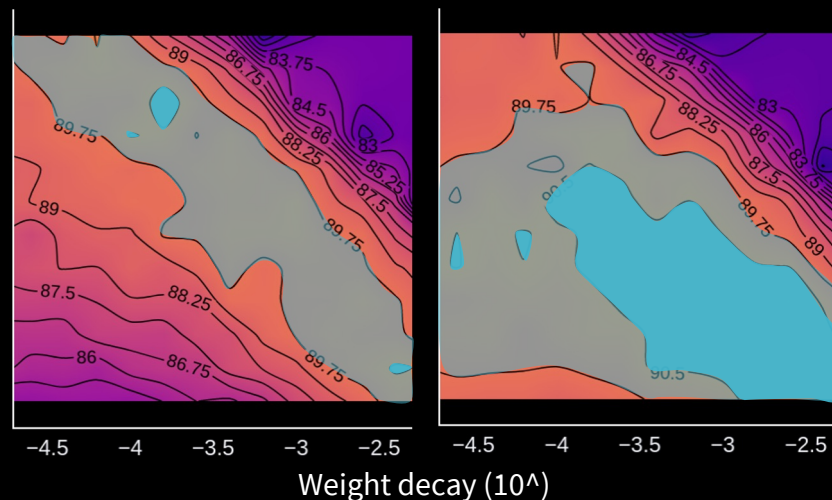
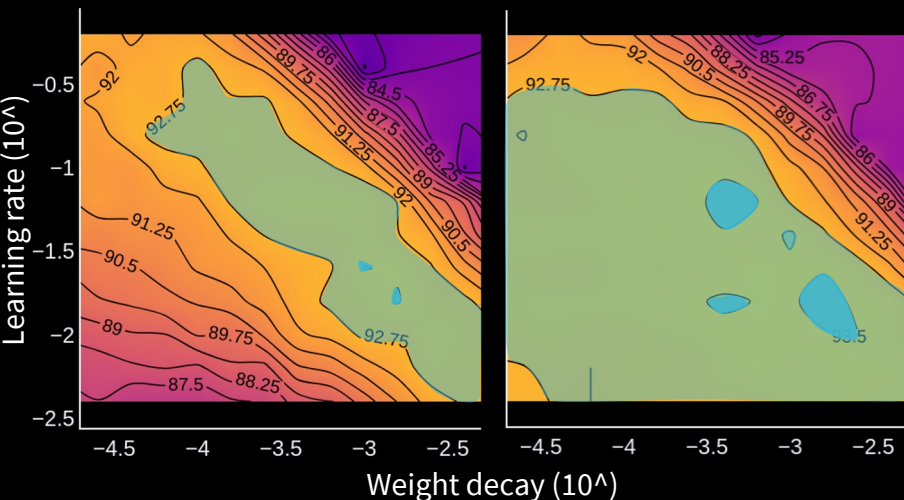
20% Non-uniform


Retraining


Fine-tuning


Retraining


Fine-tuning



 > 93.50% acc

 > 90.50% acc

 > 92.75% acc

 > 89.75% acc

Magnitude Pruning

embed

# MobileNetV2 on CIFAR-10

60% Non-uniform

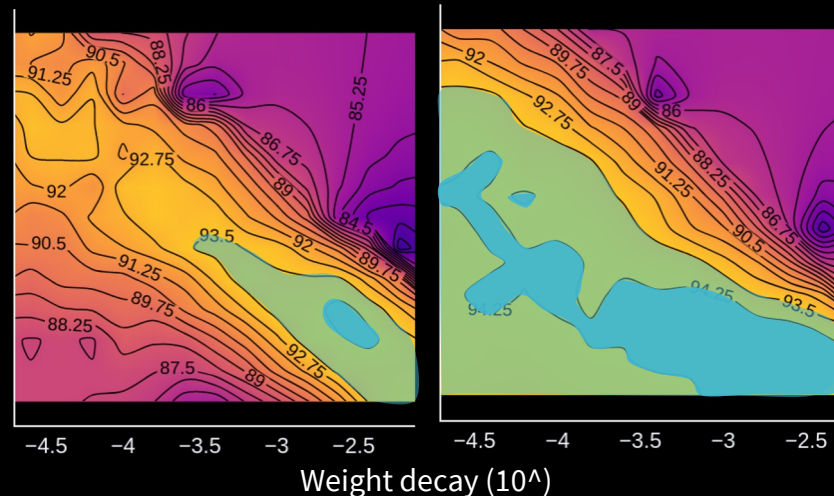
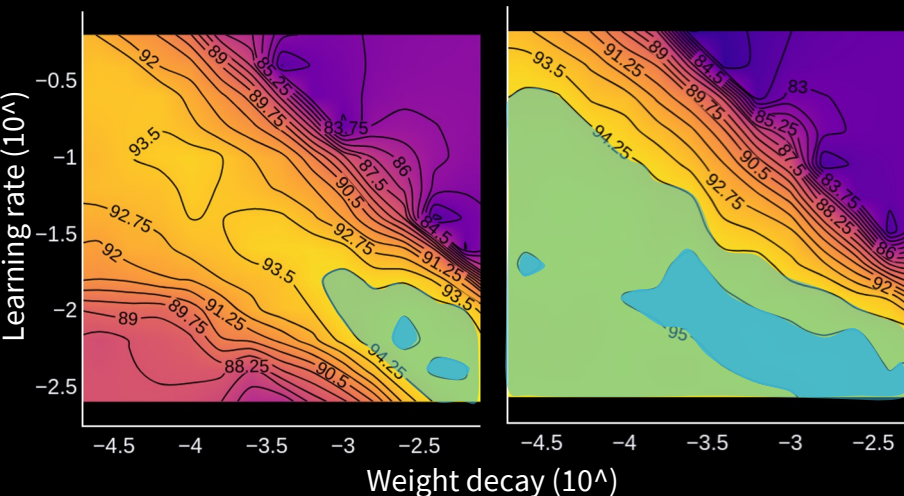
40% Non-uniform

Retraining

Fine-tuning

Retraining

Fine-tuning



Magnitude Pruning

embed

# Are Hyperparameters Overrated?

From large models to tinyML

What have we seen today?

Importance of Pruning for  
deployment on the edge

Experiments on the impact of  
structured pruning on learning rate  
and weight decay

## Conclusions

Pruning (retraining from scratch) does not have a significant impact on the overall shape and structure of the WD-LR space

- *Optimal hyperparameters of the basemodel can serve as a reasonable starting point for the pruned model*
- *When fine-tuning we are more likely to fall within an optimal area*
  - *model is more robust to HP changes*



# Are Hyperparameters Overrated?

From large models to tinyML

What have we seen today?

Importance of Pruning for  
deployment on the edge

Experiments on the impact of  
structured pruning on learning rate  
and weight decay

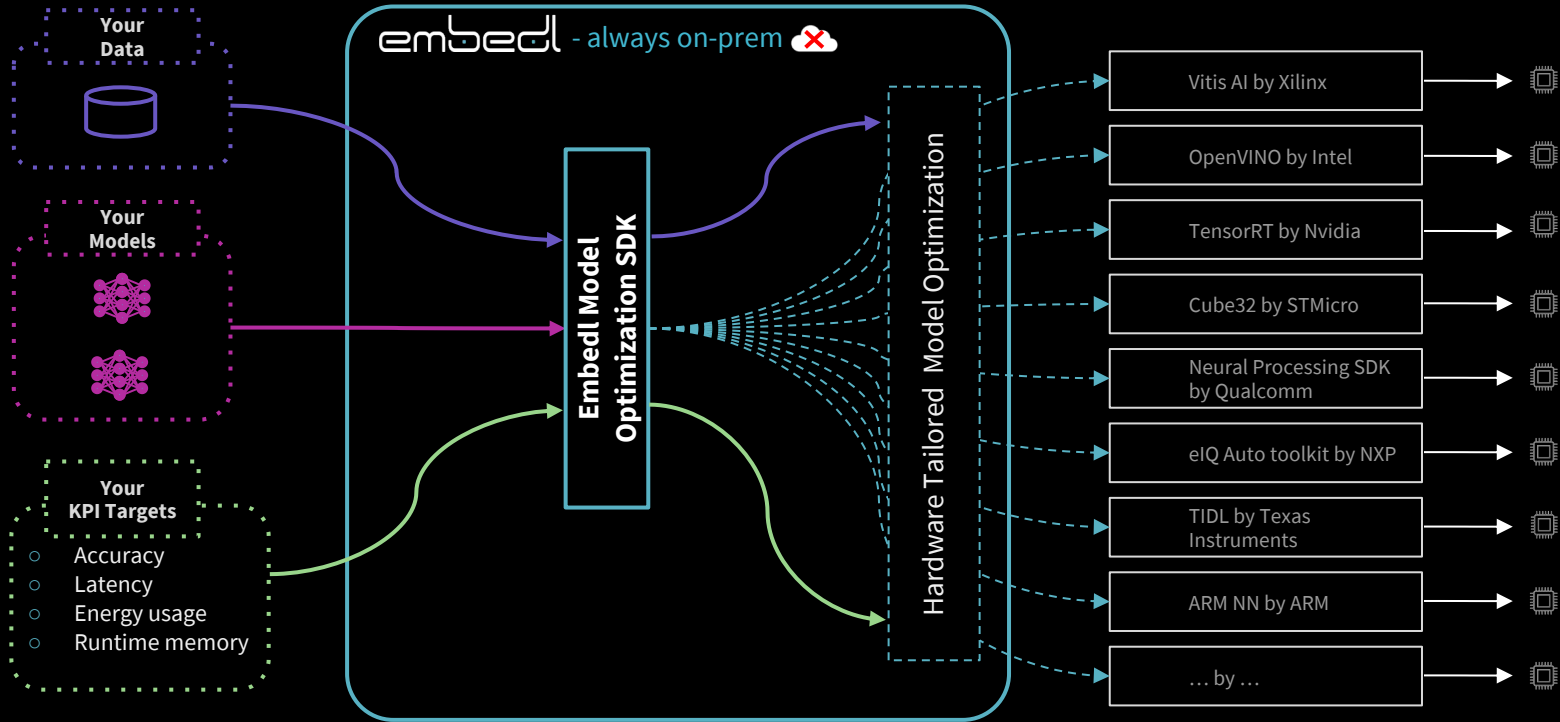
## Conclusions

Pruning (retraining from scratch) **does not have a significant impact** on the overall shape and structure of the WD-LR space

- *Optimal hyperparameters of the basemodel can serve as a reasonable starting point for the pruned model*
- *When fine-tuning we are more likely to fall within an optimal area*
  - *model is more robust to HP changes*



# Our Solution - Efficiency & Flexibility



# Copyright Notice



This presentation in this publication was presented as a tinyML® EMEA Innovation Forum. 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.tinyml.org](http://www.tinyml.org)**