# Memory-Efficient Neural Architectures

Julien Farah

2024-09-09

# Efficiency Methods

- Model pruning
- Weight quantization

# Advantages

- Reduced inference time
- Lower energy footprint