THOR: A Generic Energy Estimation Approach for On-Device Training

Jiaru Zhang[1] Zesong Wang[1] Hao Wang[2] Tao Song[1] Huai-an Su[3] Rui Chen[3]
Yang Hua[4] Xiangwei Zhou[5] Ruhui Ma[1] Miao Pan[3] Haibing Guan[1]

[1]Shanghai Jiao Tong University [2]Stevens Institute of Technology
[3]University of Houston [4]Queen's University Belfast [5]Louisiana State University

2025.6

Motivation

- Estimating DNN training's energy cost is important.
- Challenges:
  - System Heterogeneity
  - Model Diversity
  - Runtime Complexity
- THOR: **partition** the entire model into layers and estimate overall energy by **layer-wise additivity property**.

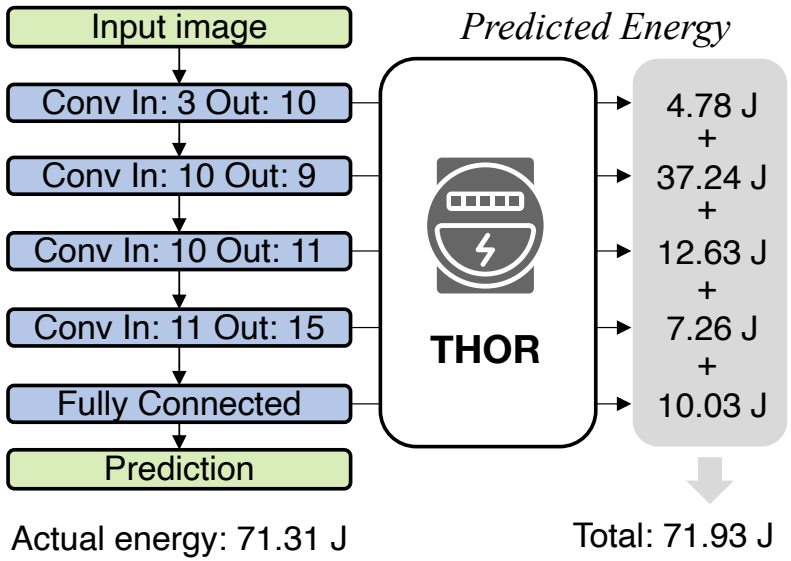
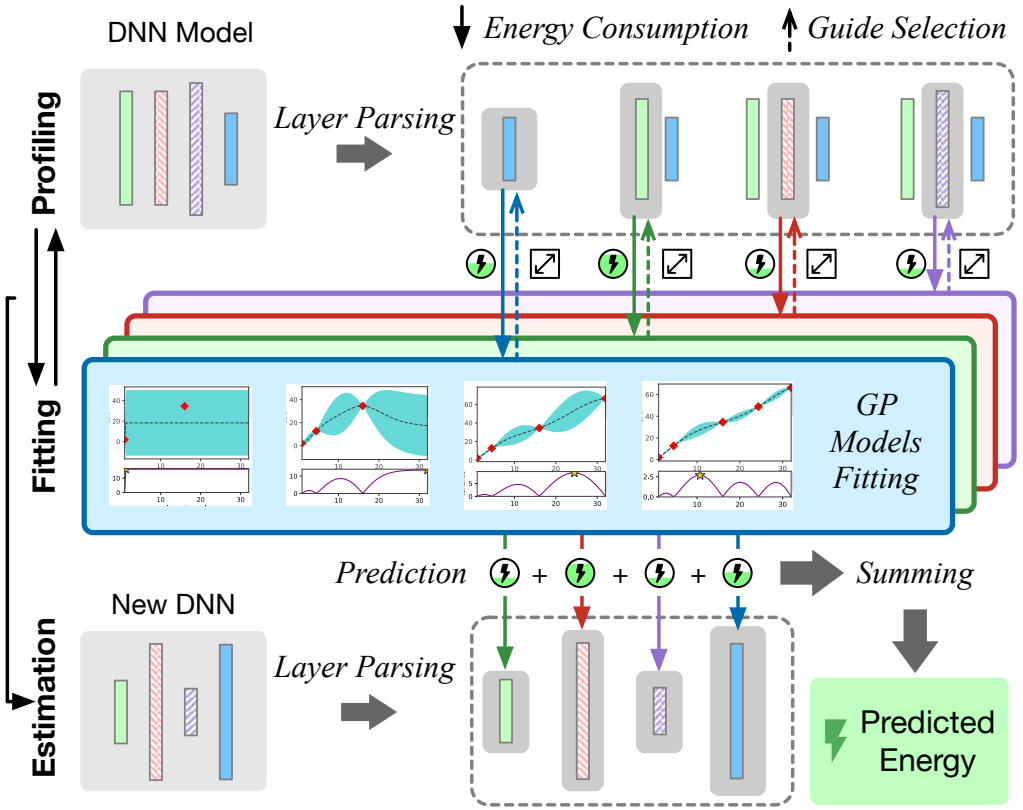


Fig: Illustration of of THOR.

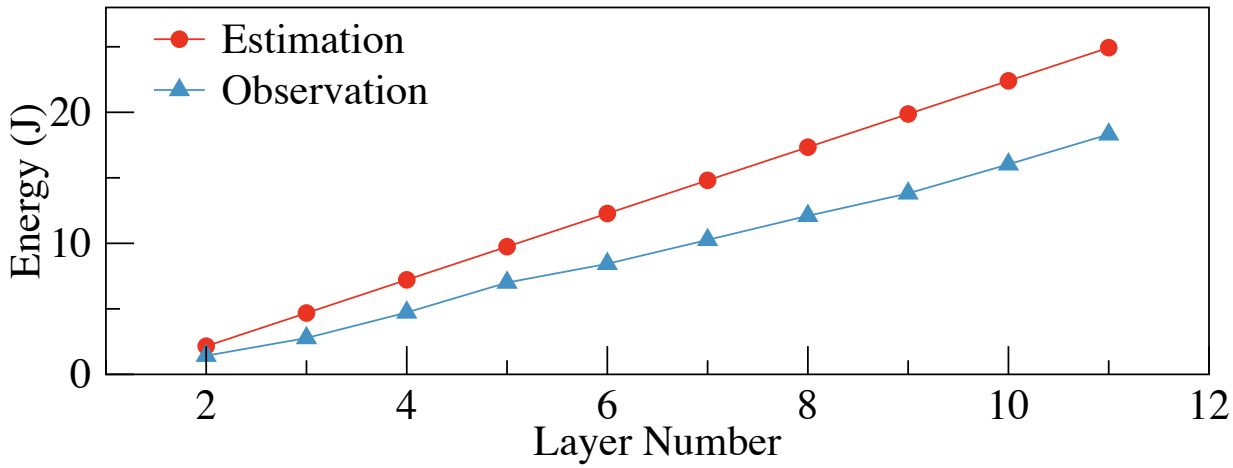
Overview of THOR

- **Profiling**: THOR partitions the DNN model into input layer, hidden layer, and output layer.
- **Fitting**: THOR separates the model as different layers and actively fits Gaussian Process (GP) models by observed layer-wise additivity.
- **Estimation**: After profiling and fitting, THOR can obtain the whole energy estimation by summing the estimated energy of each layer.



Profiling

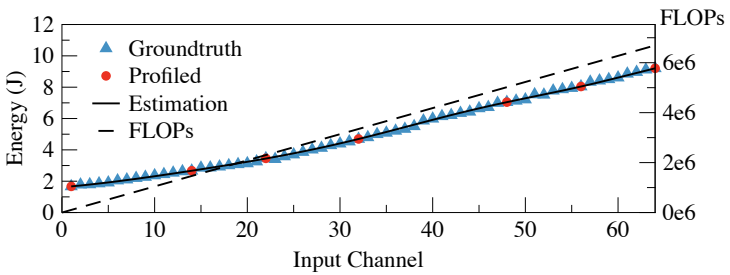
- Observation: Layer-wise Energy Additivity of DNNs

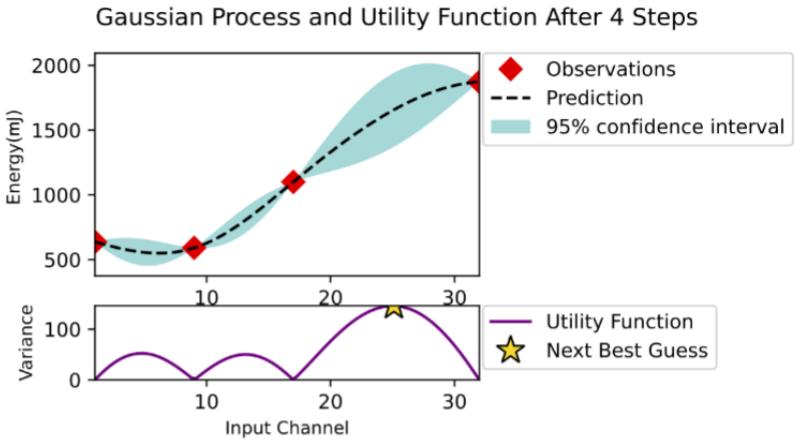


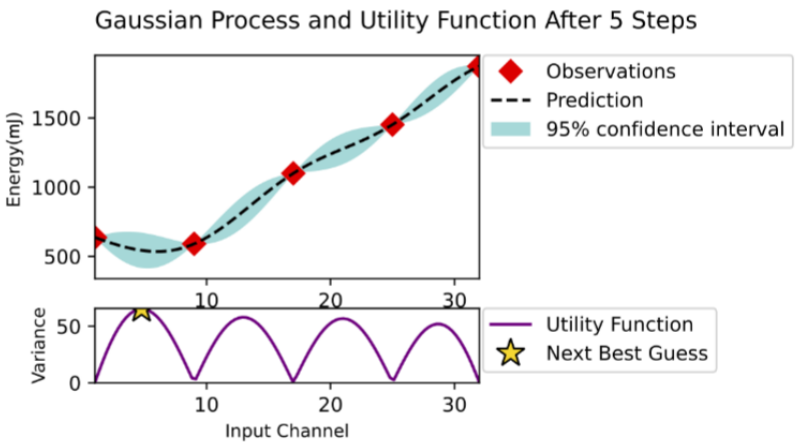
- Profiling process:
  1. Firstly, profiling the output layer by treating it as a single-layer model
  2. Secondly, profiling the input layer by subtracting the output layer's costs from a two-layer model
  3. Finally, profiling the middle layer by subtracting the input and output layers' costs from a three-layer model

Fitting

- THOR utilizes GP model to fit energy consumption characteristics of layers.
- GP models guide the selection of the next point.
- GP models are more accurate compared with FLOPs-based estimation.







Estimation

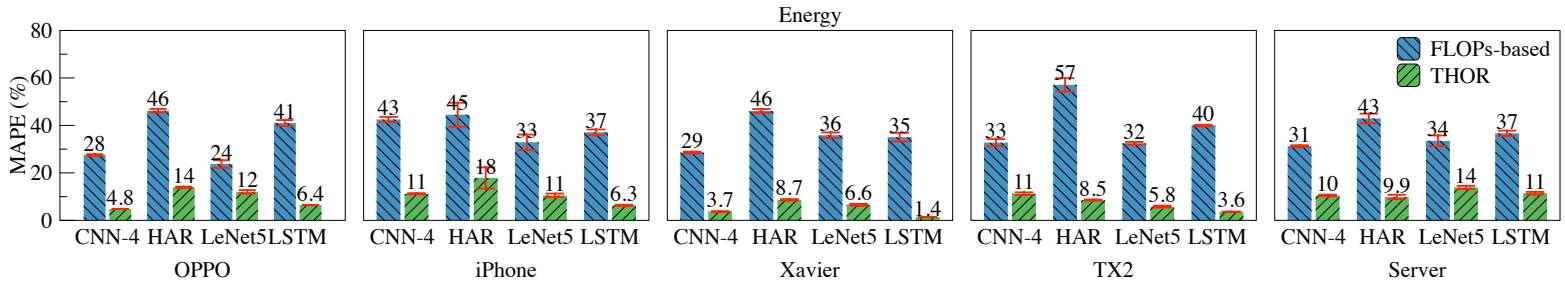
- After profiling and fitting, the total energy can be estimated by summing the estimated energy of all layers.

$$\hat{E}_{model} = \hat{E}_{input}(C_1) + \sum_{i=2}^{n-1} \hat{E}_{hidden}(C_{i-1}, C_i) + \hat{E}_{output}(C_{n-1}) \qquad (1)$$

- It performs well under the system heterogeneity and model diversity by fitting separately.

End-to-End Estimation Evaluation

- THOR outperforms FLOPs-based evaluation across all five devices and four networks.



- THOR profiling are usually completed within 20 minutes.
- THOR is effective on larger networks like ResNets and Transformers.

Table 1: Time cost (sec) of profiling and fitting.

	LeNet5	5-layer CNN	HAR	LSTM
OPPO	694	1688	2188	1615
iPhone	1201	1012	2446	1168
Xavier	184	421	740	1145
TX2	285	1211	4433	422
Server	235	268	562	436

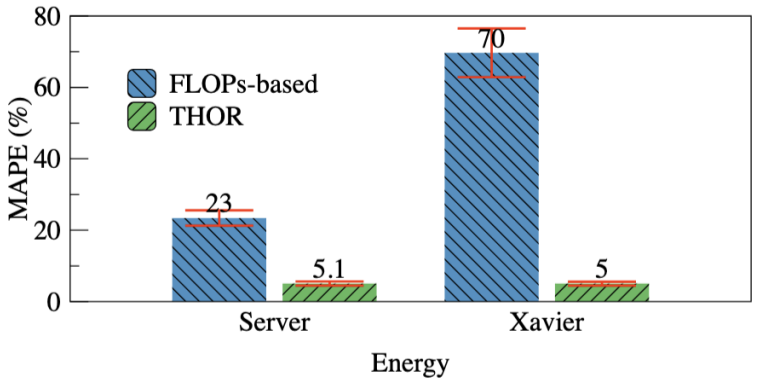
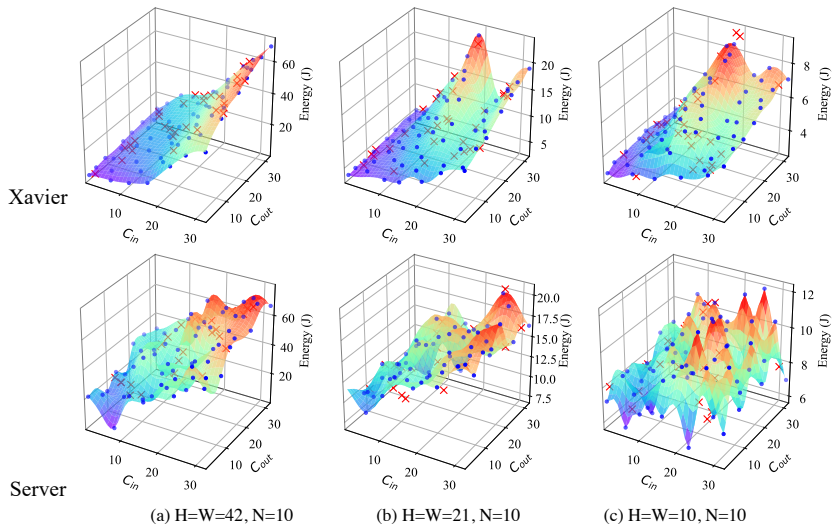
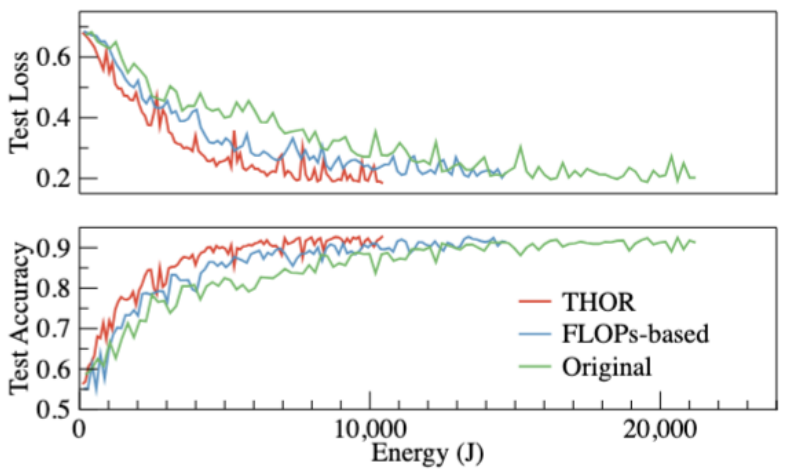


Figure 9: Energy estimation of Transformer.

Layer Characteristics and Case Study

- THOR can still accurately estimate the energy costs under complex layer characteristics.
- THOR can guide energy-conscious model pruning to create a leaner architecture with the same performance and 50% energy consumption.





Conclusion

- This paper proposes **THOR**, a generic method to estimate the energy consumption of DNN training.
- **GP** is used to fit layerwise consumptions, then the end-to-end estimation can be obtained by summing the energy consumption predictions of each layer based on the presented **layer-wise energy additivity**.
- THOR is **effective** across different architectures and devices. It can also be **integrated into existing training frameworks** to guide energy-aware job scheduling.