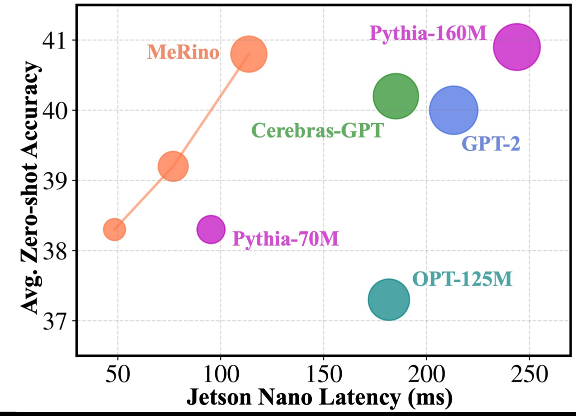
MeRino: Entropy-Driven Design for Generative Language Models on IoT Devices

Youpeng Zhao, Ming Lin, Huadong Tang, Qiang Wu, Jun Wang

Motivation:

1. Deploying large language models (LLMs) on cloud computing platforms are expensive (energy consumption/financial costs)
2. Edge computing makes deployment of LLMs on resource-constrained devices an appealing solution to promote sustainability, accessibility and privacy
3. Integration of LLMs into mobile devices are challenging
   a. Existing LLMs are costly to deploy (memory footprint/latency)
   b. Hardware configurations are heterogeneous



Methodology:

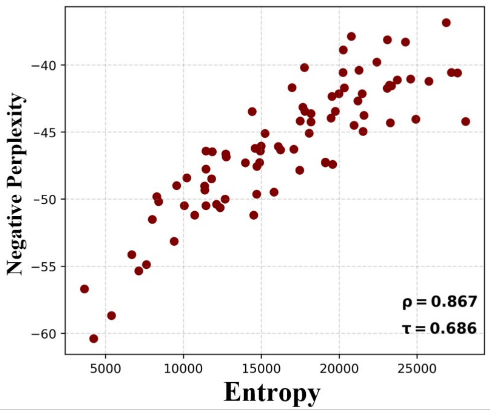
Predicting LLM performance using entropy (without training)

- **Definition of entropy:** $\widehat{H}(W) \triangleq \mathbb{E}\{\sum_{j=1}^{r_i} \log(1+\frac{s_j^2}{\epsilon^2})\}$
- **Depth-width ratio:** $\gamma = \beta L/\hat{w}$
- **Transformer entropy:** $\hat{w}_{\text{MHSA}} = \log E \quad \hat{w}_{\text{FFN}} = \log F$

$$\widehat{H}_{\text{MHSA}} = (1-\frac{\beta L}{\hat{w}_{\text{MHSA}}})\sum_{i=1}^{L}\widehat{H}(W_i^Q, W_i^K, W_i^V, W_i^O)$$

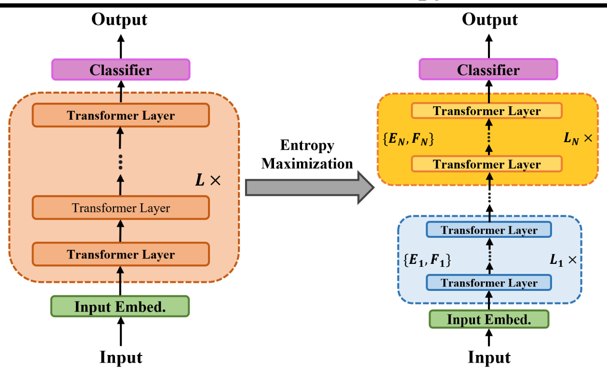
$$\widehat{H}_{\text{FFN}} = (1-\frac{\beta L}{\hat{w}_{\text{FFN}}})\sum_{i=1}^{L}\widehat{H}(W_i^{\text{FFN}_1}, W_i^{\text{FFN}_2})$$

$$\widehat{H} = \alpha_1 \widehat{H}_{\text{MHSA}} + \alpha_2 \widehat{H}_{\text{FFN}}$$



Model Design:

1. Block-wise parameter sharing
   Memory footprint reduction
2. Constrained search space
   187K architectures
3. Evolutionary search



Results:

Method	Search Device	Search Time (h)	Energy Costs (Wh)	Average Acc.
TE-NAS	GPU*	1.2	300	0.389
Ours	CPU†	0.05	0.75	**0.408**

	FLOPs (↓)	Latency (↓)	**WikiText-2**	**PTB**
Pythia-70M	100 G	95 ms	40.95	60.28
Pythia-162M	270 G	243 ms	23.52	36.02
Cerebras-111M	260 G	185 ms	36.93	51.89
GPT-2-124M	290 G	213 ms	25.19	33.95
OPT-125M	210 G	182 ms	23.62	29.02
OPT-350M	720 G	559 ms	**18.51**	**23.08**
MeRino-52M	60 G	48 ms	39.05	52.18
MeRino-61M	110 G	77 ms	34.24	34.11
MeRino-64M	160 G	114 ms	22.47	27.06