Pruna AI

Compression for Language Models

Bertrand Charpentier
Founder, President & Chief Scientist

How Do People Use Deep Learning?

Many Deep Learning Tasks

Image
- Classification
- Object detection
- Gen. from text
- Segmentation
- …

Video
- Object tracking
- Gen. from text
- Gen. from image
- …

Audio
- Transcription
- Translation
- Gen. from text
- …

Text
- Question answering
- Summarization
- Classification
- …

Proteins
- Folding
- De nuovo Gen
- Property prediction
- …

Phases of Deep Learning Models

80-90% of DL workload

Development → Training → Inference

[1] The efficiency misnomer, ICLR 2022

Why Do We Need Efficient Deep Learning?

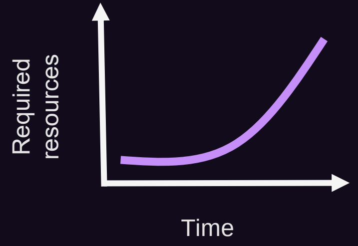
Direct Costs	Money	Time	Memory	Energy/CO_2
Critical Use Cases	Budget constraints	User experience Real-time reaction	Edge portability Data privacy	ESG considerations

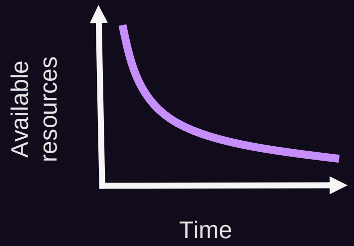
What Are Good Efficiency Metrics?

Metrics are often correlated

Metrics can be contradictory

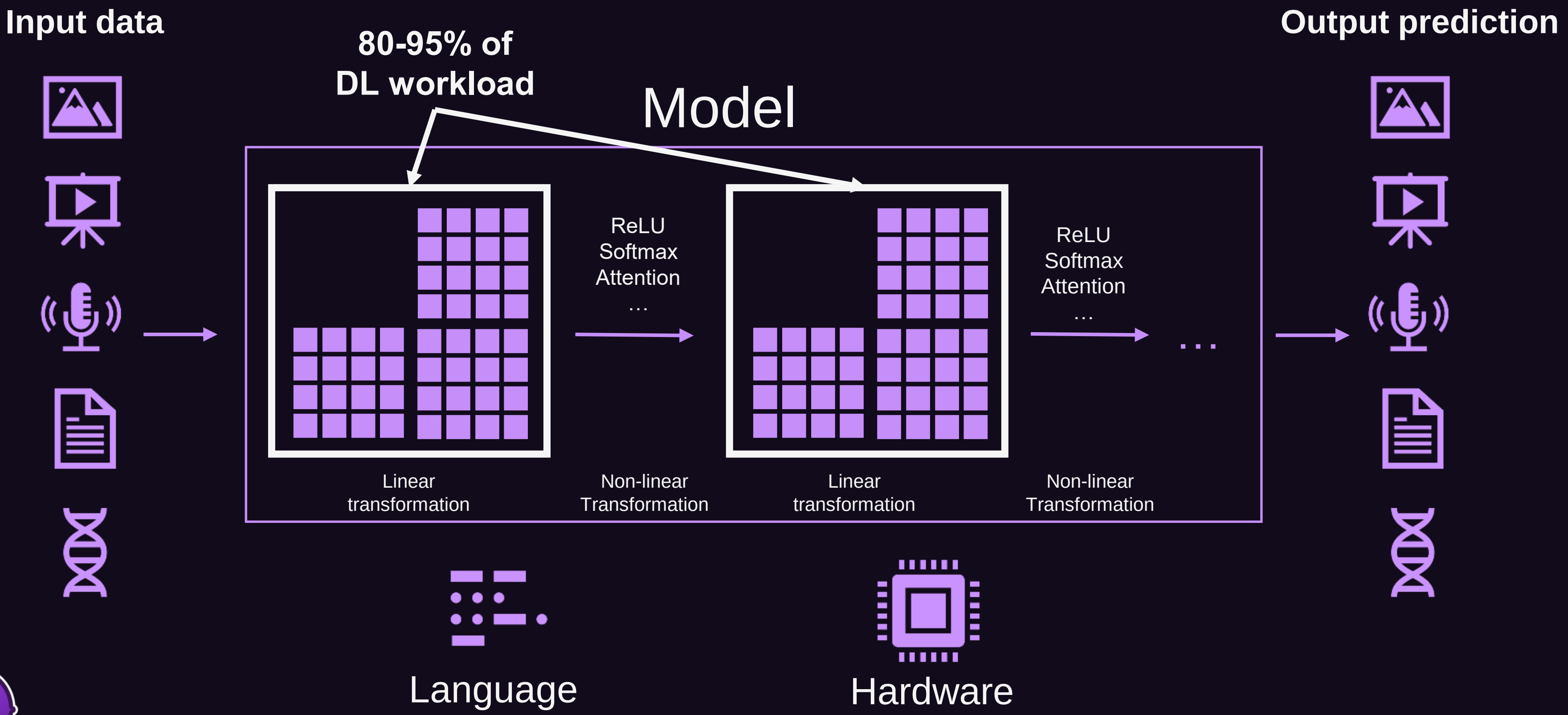
Metrics should measure **direct real-world costs**





[1] The efficiency misnomer, ICLR 2022
[2] Power Hungry Processing: Watts Driving the Cost of AI Deployment?

How Does a Deep Learning Model Work?

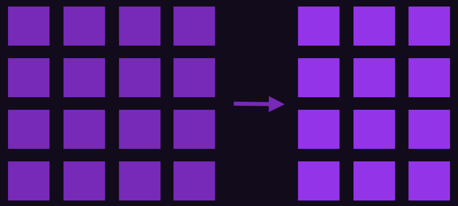


[1] Accuracy is not the only Metric that matters: Estimating the Energy Consumption of Deep Learning Models, TCC - ICLR 2023

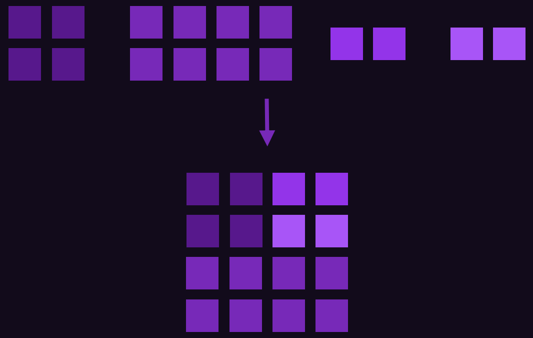
Pruning

Quantization

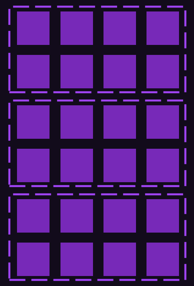
Distillation



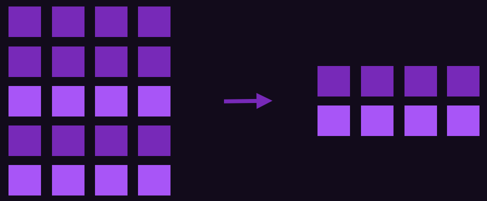
Compilation



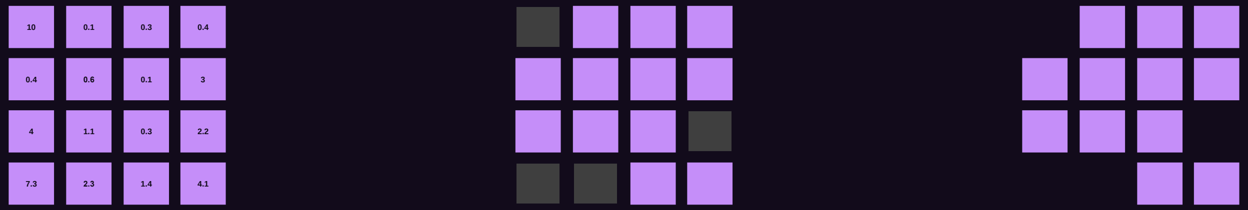
Batching



Caching



How to Prune Deep Learning Models?



Step 1
Score structures

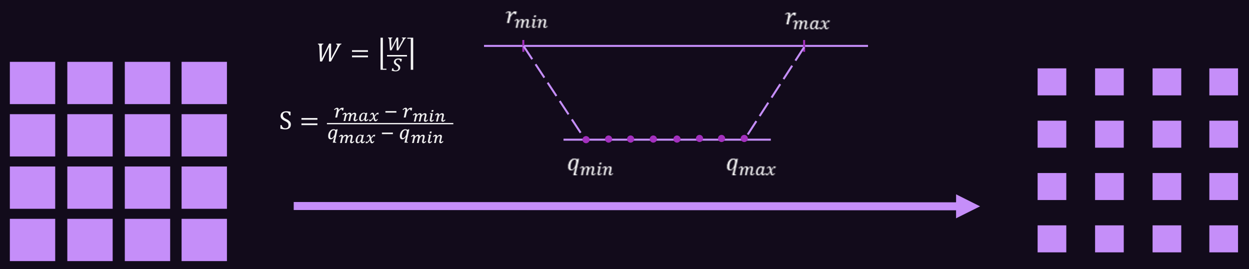
Step 2
Rank structures w.r.t. scores

Step 3
Prune structures with lowest scores

- **What structure to prune?** Unstructured pruning, structured pruning, …
- **How to score structures?** Random, magnitude, gradient, hessian
- **What sparsity to prune?** Homogeneous, heterogreneous
- **When to prune?** Before, during, after training

[1] Winning the Lottery Ahead of Time: Efficient Early Network Pruning. ICML 2022
[2] How Sparse Can We Prune A Deep Network: A Fundamental Limit Viewpoint
[3] Structurally Prune Anything: Any Architecture, Any Framework, Any Time.

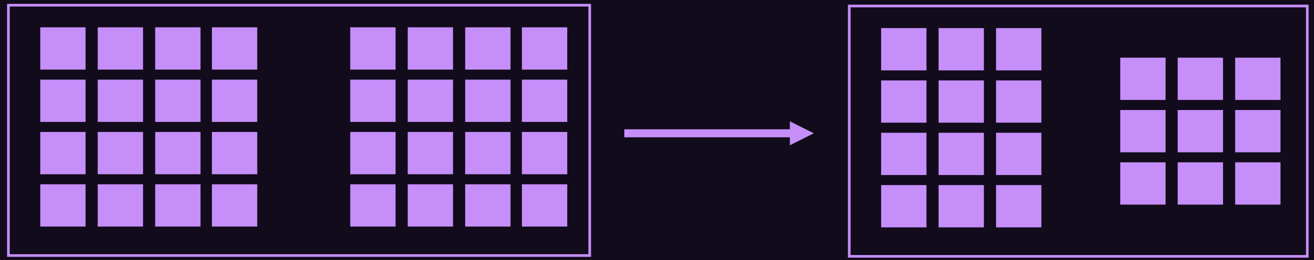
How to Quantize Deep Learning Models?



- **What structure to quantize?** Per tensor/channel/group/outliers, weight/activation
- **How to quantize structures?** Linear quantization, code books
- **What precision to quantize?** 16, 8, 4, 2, 1 bits
- **When to quantize?** Quantization-aware, post-training

[1] AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. NeurIPS 2023
[2] GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. ICLR 2023

How to Distill Deep Learning Models?



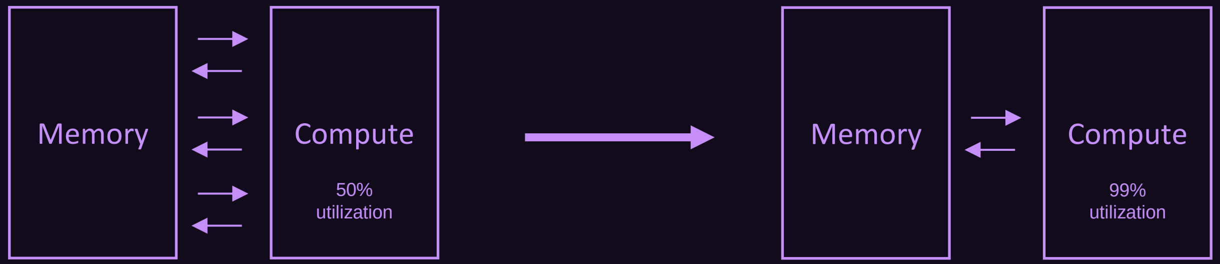
Teacher model

Student model

- **What information to distill?** Response, feature, weights…
- **What model to distill into?** Architecture, size, precision
- **When to distill?** Offline, online

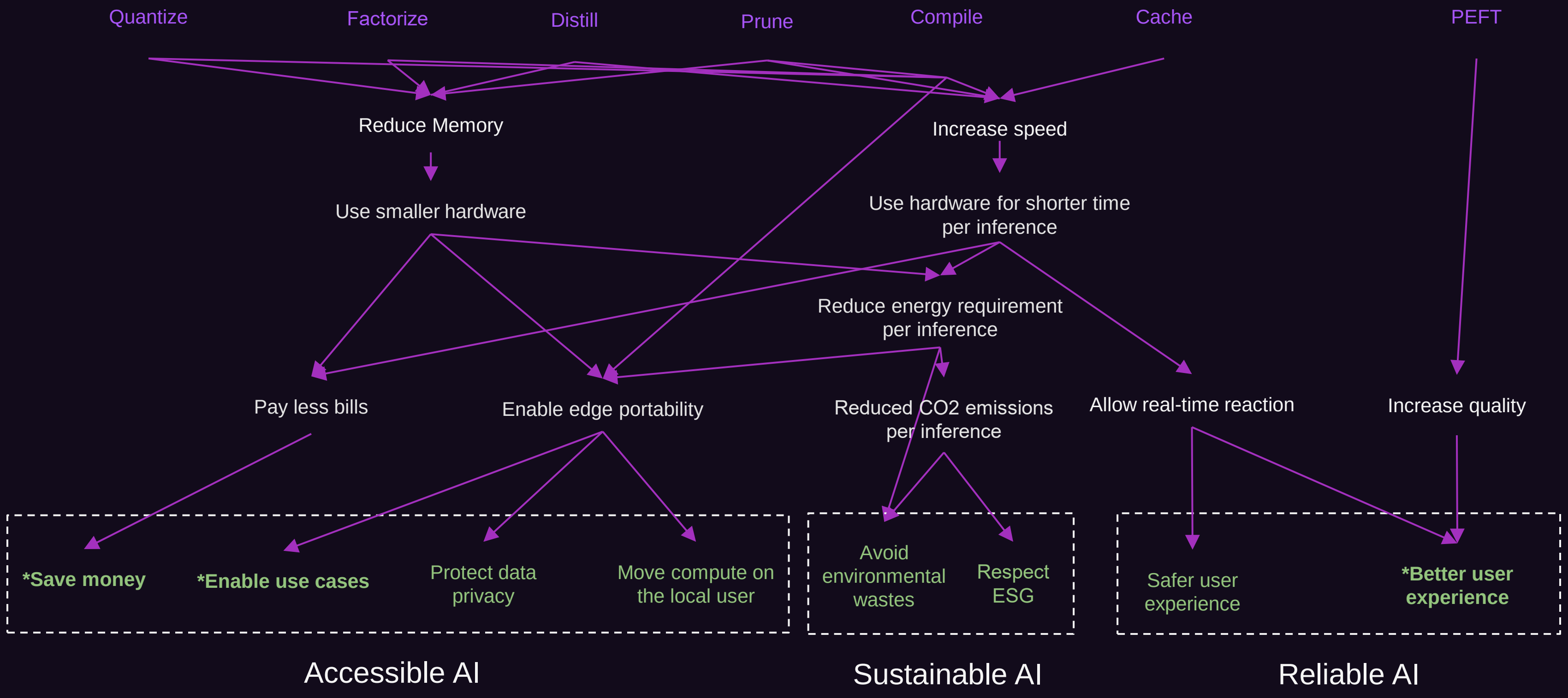
[1] Does Knowledge Distillation Really Work?. NeurIPS 2021
[2] Improved Knowledge Distillation via Teacher Assistant. AAAI 2019
[3] Fitnets: Hints for Thin Deep Nets. ICLR 2015

How to Compile Deep Learning Models?



- **What structure to compile?** Linear/Attention/…**,** Fuse operators
- **What compilation backend/kernels to use?** CUDA, Triton, ARM, Custom backend
- **What hardware is supported by compilation?** CPU, GPU, others
- **How to compile?** Memory vs compute bound

How Does Compression Benefit Deep Learning?



How Well Does Compression Methods Work?

	Acc.	Speed	Mem.
Base	~75%	x1	x1
Prun.	~76%	x2	X0.5

ResNet50 - ImageNet

	Perpl.	Speed	Mem.
Base	~5.6	x1	x1
Quant.	~6.0	x2	X0.5

Llama 7B - WikiText

	Speed	Mem.
Base	x1	x1
Distill.	x2	X0.5

Stable Diffusion

- **Remark 1:** There are many, many, many other compression methods.
- **Remark 2:** Compression methods can be combined
- **Remark 3:** The best (combination of) compression methods depends on the final application setup (incl. architecture, hardware, data,…).

Compression of DL model is complex!

[1] Structurally Prune Anything: Any Architecture, Any Framework, Any Time.
[2] GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. ICLR 2023
[3] SSD-1B. Segmind

How Well Does Compression Methods Work?

Base

Pruna

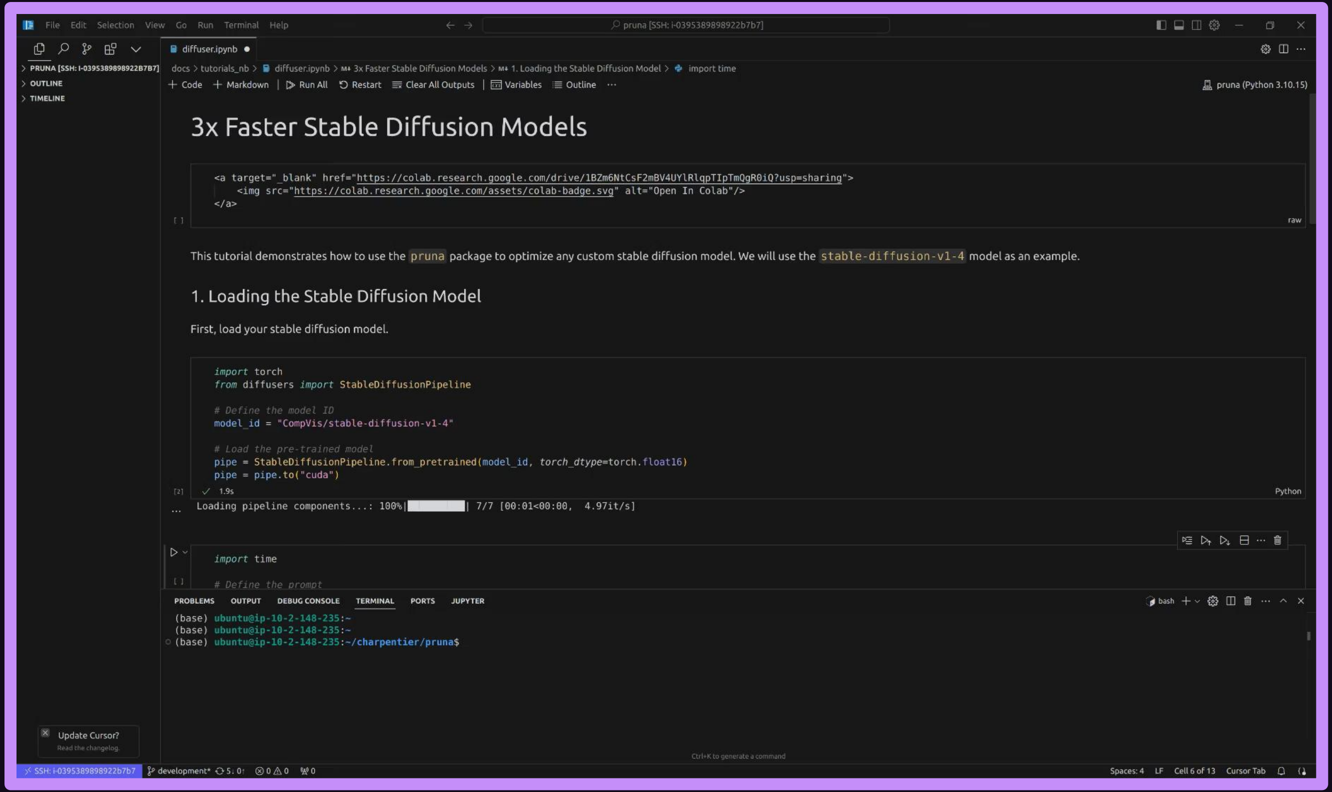
How Well Does Compression Methods Work?

Base

Pruna

Compression of DL model does not necessarily means quality loss!

How to Compress Your Deep Learning Model?



How to Compress Your Deep Learning Model?

Manual solution

1. Read & understand research papers.
2. Implement & test compression methods
3. Integrate compression methods for your specific model/hardware.
4. Test & evaluate all hyperparameters.
5. *Hopefully* get efficiency gains

Pruna AI solution

1. Install Pruna package
2. Smash your AI model
3. Get *significant* efficiency gains

Long research exploration $\xrightarrow{\text{Accessibility}}$ Easy-to-use compressions

Painful model/hardware debugging $\xrightarrow{\text{Adaptability}}$ Model/hardware adaptability

Resources wasted/Impossible project $\xrightarrow{\text{Performance}}$ Reliable efficiency gains

Try our 10,000+ smashed models on Hugging Face!