Automated Model Compression via Genetic Algorithms

Gabriel MATAR, Octave, Gabriel Mendes, Paul Plantier

Introduction and Scientific Background

- **The central issue in modern AI: How can we reduce the size of an LLM (such as GPT-2) to make it usable on standard devices without sacrificing its intelligence?**
- The Foundation (Paper 2 - Wang et al.): This paper, “When LLMs Meet Evolutionary Algorithms”, is our basis.
- Methodology (Paper 1 - OptiShear - Liu et al. & Paper 4 - EvoP - Wu et al.): They introduce the concept of heterogeneous layer sensitivity.
- Contrast (Paper 3 - Lang et al.): This paper on quantisation allows us to define our subject.
- The Critical Perspective (Paper 5 - Self-Pruner - Huang et al.): Finally, this very recent paper on self-pruning will serve as a point of comparison for discussing the limitations of our method at the end of the presentation.

Problem Definition - The Gradient Impasse

- Training a neural network is a continuous and differentiable optimisation problem.
- Gradient descent is used because the loss curve is smooth.
- However, structural pruning (choosing which neurons to keep) is a discrete and combinatorial problem.
- This is where our technical choice is justified: we use a Genetic Algorithm. This is a Global Search method.

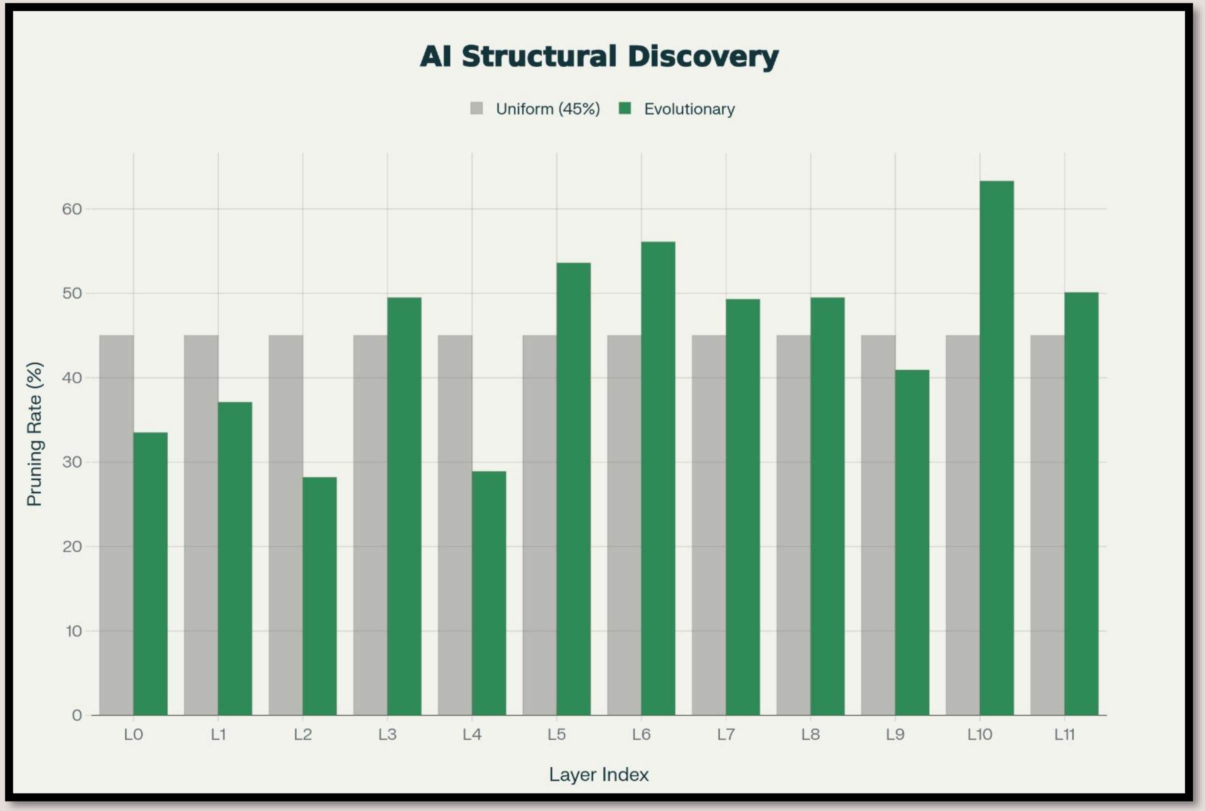
TECHNICAL APPROACH

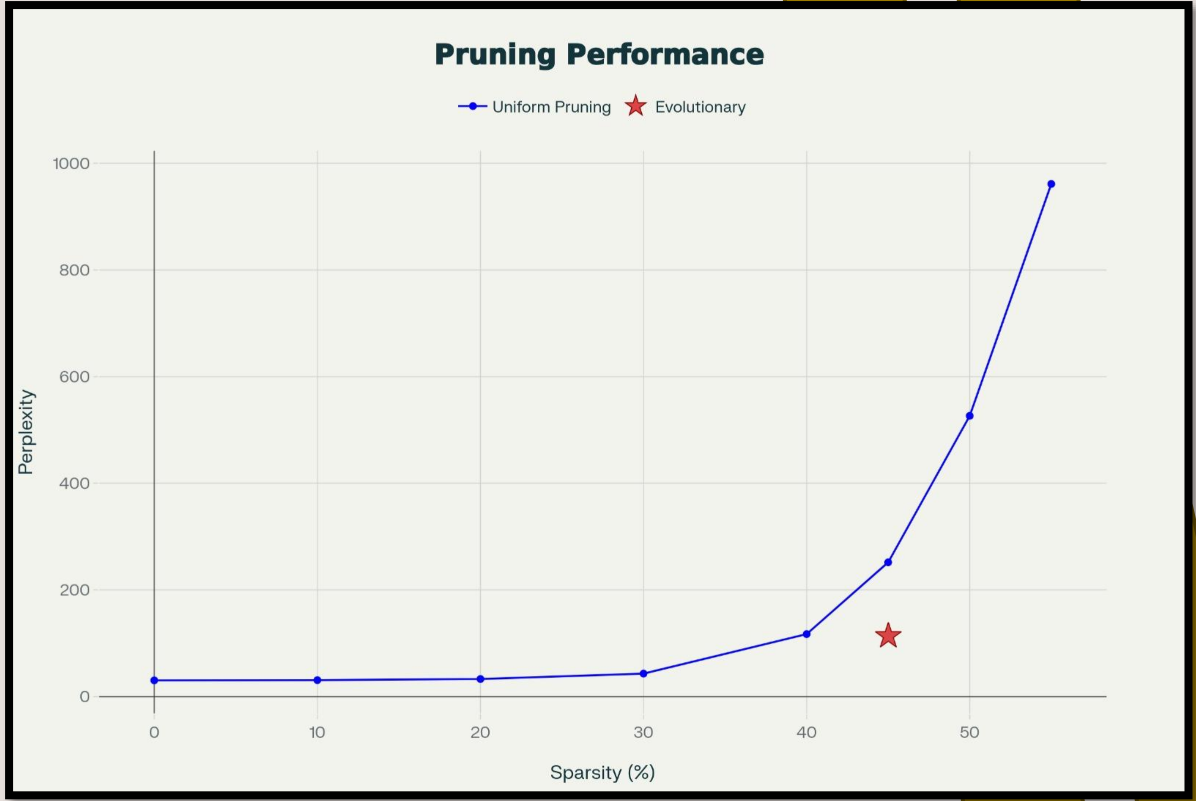
the intelligence of “layer-wise" pruning

Pillar 1 Genome Encoding	Pillar 2 Fitness Function	Pillar 3 Evolutionary Cycle
$C = [\alpha_1, \alpha_2, \ldots, \alpha_{12}]$ • 12 continuous variables • One per layer • Search space reduced 1000x	PPL = Perplexity (minimize) • Sandard LM metric • Lower = Better	Initial Pop : 20 arch ↓ [selection] [crossover] [mutation] ↓ Gen 2, 3, ..., 50 → Best genome found

apply_pruning_to_model(genome)	evaluate_genome(genome) → →PPL	run_genetic_algorithm(50 gens)
→ Modifies model architecture in real-time → Uses L1 Unstructured Pruning (PyTorch native)	→ Deepcopy model (preserves original) → Runs inference on calibration set	→ Implements elitism (Keep top 50%)

Results & Analysis





Critical Discussion, Limitations and Comparison

- Comparison with Quantization (Paper 3 - Lang et al.):

- Limits and Openness (Self-Pruner - Huang et al.):

  Genetic algorithm

  Self-Pruner

Team & Conclusion

- **[Everyone] – Research & Theory:** Analysis of papers and definition of mathematical constraints.
- **[Octave & Gabriel Ma] – Architecture & Algorithm:** Code development and Genetic Algorithm implementation.
- **[Gabriel Me & Paul] – Analysis & Interpretation:** Interpretation of results, discussion of implementation limitations, and areas for improvement.

Key Learning:

- AI optimization is not limited to the training phase.
- We experienced the constraints imposed by **layer heterogeneity**.
- We learned that neural architecture search is a **combinatorial problem**, requiring global search strategies (like Genetic Algorithms) rather than local gradient-based methods.