Semantic Thermodynamic Entropy Pruning: Key Insights

Core Principles

1. **Entropy-Guided Selection**: Pruning decisions based on information-theoretic measures rather than magnitude alone
2. **Semantic Preservation**: Maintaining meaningful content while reducing computational overhead
3. **Thermodynamic Optimization**: Applying energy minimization principles to neural network compression
4. **Information Conservation**: Preserving mutual information between network layers

Mathematical Framework

- **Landauer's Principle**: Minimum energy E â‰¥ k_B * T * ln(2) for bit erasure
- **Semantic Entropy**: H_s â‰¤ H (semantic entropy bounded by syntactic entropy)
- **Information Concentration**: IC = rank(W) * H(activations) / ||W||_0
- **Energy Efficiency**: Î· = (semantic_fidelity * compression_ratio) / energy_consumption

Performance Metrics Summary

Aspect	Traditional Pruning	Entropy-Based Pruning	Semantic-Thermodynamic Pruning
Compression	60-80%	85-95%	80-90%
Accuracy Loss	5-15%	2-10%	3-13%
Energy Savings	40-60%	70-85%	75-88%
Semantic Preservation	Low	Moderate	Very High

Key Algorithms

1. **NEPENTHE**: Entropy-based depth reduction through layer linearization
2. **MIPP**: Mutual Information Preserving Pruning
3. **Energy-Aware Pruning**: Direct energy consumption optimization
4. **Thermodynamic Filter Selection**: Temperature-based filter importance

Applications

- Large Language Model optimization
- Edge computing deployment
- Medical AI safety systems
- Real-time semantic processing
- Resource-constrained environments