


Pruning OPT for Summarization

Vinay Padegal, Sarang Sridhar, Chris Lee,
Andrew Zolensky

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Our Topic

Large Language Model Pruning Techniques...

...remove connections, by setting weights to zero or deleting (sub)layers.

...drop connections based on importance.

...aim to reduce compute, preserve performance

Formal Problem Definition

Measure the compute and summarization performance for pruned, fine-tuned LLM's.

An Example

Most LLM's are too large to deploy in embedded medical devices without incurring latency

Pruning layers in a performance-sensitive way could help models fit on smaller devices

Motivation

LLMs consume a large amount of compute for inference alone

Leads to high costs, energy consumption, carbon footprint

Larger Objectives

Reduce the carbon footprint associated with LLMs and deep learning broadly, as it becomes the norm for modern NLP applications

Demonstrate utility of data-aware pruning methods for specializing for a task while reducing model size

Enable wider deployment of LLMs

Relation to Class

Parameter-efficient fine-tuning, quantization, pruning connections and sub-layers

Pruning connections does not reduce runtime complexity, but structured pruning at the level of the neuron or layer can.

What We've Learned

Structured pruning can decrease runtime

Pruning has varying impact on task-specific performance, which can be controlled by fine-tuning and activation-aware pruning techniques.

Often possible to achieve very similar performance using only a small sub-network of the initial trained network (LTH).

Literature Themes

- Parameter quantization
- Pruning level
 - Parameter
 - Neuron
 - Layer
- Pruning + Training Approaches
 - Prune & Fine-Tune
 - Activation-based pruning
 - Quantization
 - QLoRA

Previous Work

- Summarization task with LLM's
- Layer-based pruning
- Magnitude pruning, Wanda pruning, SparseGPT
- QLoRA fine tuning

Method	Weight Update	Sparsity	LLaMA				LLaMA-2		
			7B	13B	30B	65B	7B	13B	70B
Dense	-	0%	59.99	62.59	65.38	66.97	59.71	63.03	67.08
Magnitude	✗	50%	46.94	47.61	53.83	62.74	51.14	52.85	60.93
SparseGPT	✓	50%	54.94	58.61	63.09	66.30	56.24	60.72	67.28
Wanda	✗	50%	54.21	59.33	63.60	66.67	56.24	60.83	67.03
Magnitude	✗	4:8	46.03	50.53	53.53	62.17	50.64	52.81	60.28
SparseGPT	✓	4:8	52.80	55.99	60.79	64.87	53.80	59.15	65.84
Wanda	✗	4:8	52.76	56.09	61.00	64.97	52.49	58.75	66.06
Magnitude	✗	2:4	44.73	48.00	53.16	61.28	45.58	49.89	59.95
SparseGPT	✓	2:4	50.60	53.22	58.91	62.57	50.94	54.86	63.89
Wanda	✗	2:4	48.53	52.30	59.21	62.84	48.75	55.03	64.14

Results from the initial Wanda Paper

SparseGPT

- Seeks sparsity masks M_L and updated weights W_L^* to minimize layer-wise reconstruction error.
- Exact optimization is NP-Hard.
- Solution: Separates mask selection and weight reconstruction.
- this involves costly Hessian inverse calculations
- Paper introduces a novel approximate sparse regression solver

Wanda

- Prunes based on the product of weights and input activations rather than weight magnitudes alone.
- Computes weight importance per-output, improving over layer-wide methods.
- Achieves pruning without gradient computations or weight updates, making it faster than SparseGPT.
- Achieves high zero-shot accuracy at 50% sparsity in LLaMA-65B and LLaMA-2-70B.

Dataset

For our analysis, we use the Billsum dataset from Huggingface. The dataset comprises bills passed in the US government and California State Legislature's 2015-2016 session.

- Each record in the dataset contains the bill title, the bill text and the summary (label).
- The dataset is divided into three distinct splits: training, validation and test
- Train split used for fine-tuning process, validation and test splits for benchmarking purposes.

Table 1: Number of Examples in Each Data Split

Split	Number of Examples
Training	5,000
Validation	100
Test	100

Metrics

Summarization Quality Metrics

- ROUGE (1, 2, L)
- BERTscore

Compute Metrics

- Storage memory
- Inference Time

Evaluation Framework

Summarization Performance

Measure semantic similarity between ground truth summaries and model outputs by using n-gram overlap metrics like ROUGE and embedding-based similarity metrics like BERTscore.

Efficiency

Measure efficiency using inference time, num parameters, and disk usage.

Simple Baseline

- Opt 125m, 0-shot
 - Opt 350m, 0-shot
-
- 125m repeat, 350m blanks
 - QLoRa slower, more params, quantized
 - FFT -> best performance on all metrics

Strong Baseline

- Opt 125m, Full Fine-Tune
- Opt 350m, Full Fine-Tune
- Opt 125m, QLoRa Fine-Tune
- Opt 350m, QLoRa Fine-Tune

Table 3: Performance of Sparse Models Across Different Architectures

Model	Technique	Sparsity	Validation Data		Test Data		Time	Nonzero Parameters	Size (MB)
			Rouge	Bertscore	Rouge	Bertscore			
OPT 125m (Base)	-	-	0.1652	0.7243	0.0782	0.3627	278.42	125,239,296	955.66
OPT 125m (Full)	-	-	0.2338	0.8391	0.4239	0.8863	262.3	125,239,296	955.54
OPT 125m (QLoRa)	-	-	0.2274	0.8079	0.3729	0.8593	1073.65	126,124,032	484.64
OPT 350m (Base)	-	-	0.0161	0.09	0.0127	0.0717	477.11	331,196,416	2526.92
OPT 350m (Full)	-	-	0.2602	0.8391	0.4158	0.8841	473.08	331,196,416	2526.92
OPT 350m (QLoRa)	-	-	0.2298	0.8413	0.397	0.879	2273.01	333,555,712	1281.69

Fully Fine-tuned

OPT-125m: "Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act - Amends the Omnibus Crime Control and Safe Streets Act of 1968 to require an eligible dependent of a public safety officer to be eligible for assistance if such officer: (1) dies in the line of duty on or after January 1, 1978; or (2) becomes permanently and totally disabled as the direct result of a catastrophic injury sustained in the line of duty on or after January 1, 1978.

Requires the applicant's spouse to be eligible for such assistance if the applicant's spouse dies or becomes permanently and totally disabled as the result of a catastrophic injury sustained in"

OPT-350m: "Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act - Amends the Omnibus Crime Control and Safe Streets Act of 1968 to revise the eligibility requirements for educational assistance to include dependent children of a law enforcement officer who dies or becomes permanently disabled as the direct result of a catastrophic injury sustained in the line of duty on or after January 1, 1978.

Allows eligible children of law enforcement officers who die or become permanently disabled as the direct result of a catastrophic injury sustained in the line of duty on or after January 1, 1978, to: (1) receive assistance under the Omnibus Crime Control and Safe Streets Act of 1968"

QLoRA Fine-tuned

OPT-125m: "Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act - Amends the Omnibus Crime Control and Safe Streets Act of 1968 to require an eligible dependent of a public safety officer to be eligible for assistance under this subpart if such officer dies in the line of duty on or after January 1, 1978."

OPT-350m: "Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act - Amends the Omnibus Crime Control and Safe Streets Act of 1968 to: (1) extend the eligibility of a public safety officer to receive benefits under this Act; (2) extend the eligibility of a public safety officer to receive benefits under this Act if the officer dies in the line of duty on or after January 1, 1978; and (3) extend the eligibility of a public safety officer to receive benefits under this Act if the officer becomes permanently and totally disabled as the direct result of a catastrophic injury sustained in the line of duty on or after January 1, 1978."

Extensions (Pruning)

- Unstructured > structured
- Wanda > Magnitude, SparseGPT
- 4:8 > 2:4
- No decrease in num params, disk size, time

- Opt 125m, Wanda, Magnitude, SparseGPT
- Opt 350m, Wanda, Magnitude, SparseGPT
- 2:4 and 4:8 sparsity
- QLoRa

Model	Technique	Sparsity	Validation Data		Test Data		Time	Nonzero Parameters	Size (MB)
			Rouge	Bertscore	Rouge	Bertscore			
OPT 125m (QLoRa)	Wanda	50%	0.1853	0.8232	0.3263	0.8616	1064.84	63,504,384	484.64
OPT 125m (QLoRa)	Wanda	4:8	0.2153	0.8263	0.2868	0.8449	1045.93	63,504,384	484.64
OPT 125m (QLoRa)	Wanda	2:4	0.2096	0.8107	0.1384	0.6088	1038.63	63,504,384	484.64
OPT 125m (QLoRa)	Magnitude	2:4	0.109	0.7771	0.2426	0.8203	1055.93	63,504,384	484.64
OPT 125m (QLoRa)	SparseGPT	2:4	0.1446	0.6918	0.1933	0.6528	1050.35	63,504,384	484.64
OPT 350m (QLoRa)	Wanda	2:4	0.1415	0.791	0.2129	0.8141	2353.38	167,957,504	1281.69
OPT 350m (QLoRa)	Magnitude	2:4	0.0391	0.7814	0.0551	0.6537	2189.13	167,957,504	1281.69
OPT 350m (QLoRa)	SparseGPT	2:4	0.2395	0.8314	0.2994	0.8467	2271.95	167,957,504	1281.69

Wanda Sparsities (OPT-125m)

Wanda 2:4: "SECTION 1. SPECIAL Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act.--The

Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act

(SOP) is amended by striking "education" and all that follows through the period at the

end and inserting "education."

SEC. 2. SPECIAL Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act.--"

Wanda 4:8: "Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act - Amends the Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act to include a public safety officer who is eligible to receive benefits under this subpart if such officer dies in the line of duty on or after January 1, 1978; an"

Wanda unstructured: "Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act - Amends the Omnibus Crime Control and Safe Streets Act of 1968 to include an eligible dependent of a public safety officer who is eligible to receive benefits under this subpart if such officer dies in the line of duty on or after January 1, 1978; and"

Magnitude and Sparsegpt (OPT-125m)

Magnitude 2:4: "Section 1217 of the Omnibus Crime Control and Safe Streets Act of 1968 (42 U.S.C. 3796d-6) is amended by adding at the end the following new

paragraph: "(b) any public safety officer who is eligible to receive

benefits under section 1201(a) or 1201(b); or

"(C) the child of any public safety officer who is"

Sparsegpt 2:4: "Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act.

This Act may be cited as the "Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act".

OPT-125m Sparsegpt 2:4

"Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act.

This Act may be cited as the ``Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act".

OPT-350m Sparsegpt 2:4

"Amends the Omnibus Crime Control and Safe Streets Act of 1968 to provide for the replacement of the "durable" and "automobile" benefits for public safety officers who are disabled.

Amends the Children's Health Insurance Program (CHIP) to provide for the replacement of the "child" and "wife" benefits for public safety officers who are disabled.

Amends the Children's Defense Fund Act of 1996 to provide for the replacement of the "child" and "wife" benefits for public safety officers who are disabled.

Amends the Children's Defense Fund Act of 1996 to provide for the"

All Results

Table 3: Performance of Sparse Models Across Different Architectures

Model	Technique	Sparsity	Validation Data		Test Data		Time	Nonzero Parameters	Size (MB)
			Rouge	Bertscore	Rouge	Bertscore			
OPT 125m (Base)	-	-	0.1652	0.7243	0.0782	0.3627	278.42	125,239,296	955.66
OPT 125m (Full)	-	-	0.2338	0.8391	0.4239	0.8863	262.3	125,239,296	955.54
OPT 125m (QLoRa)	-	-	0.2274	0.8079	0.3729	0.8593	1073.65	126,124,032	484.64
OPT 125m (QLoRa)	Wanda	50%	0.1853	0.8232	0.3263	0.8616	1064.84	63,504,384	484.64
OPT 125m (QLoRa)	Wanda	4:8	0.2153	0.8263	0.2868	0.8449	1045.93	63,504,384	484.64
OPT 125m (QLoRa)	Wanda	2:4	0.2096	0.8107	0.1384	0.6088	1038.63	63,504,384	484.64
OPT 125m (QLoRa)	Magnitude	2:4	0.109	0.7771	0.2426	0.8203	1055.93	63,504,384	484.64
OPT 125m (QLoRa)	SparseGPT	2:4	0.1446	0.6918	0.1933	0.6528	1050.35	63,504,384	484.64
OPT 350m (Base)	-	-	0.0161	0.09	0.0127	0.0717	477.11	331,196,416	2526.92
OPT 350m (Full)	-	-	0.2602	0.8391	0.4158	0.8841	473.08	331,196,416	2526.92
OPT 350m (QLoRa)	-	-	0.2298	0.8413	0.397	0.879	2273.01	333,555,712	1281.69
OPT 350m (QLoRa)	Wanda	2:4	0.1415	0.791	0.2129	0.8141	2353.38	167,957,504	1281.69
OPT 350m (QLoRa)	Magnitude	2:4	0.0391	0.7814	0.0551	0.6537	2189.13	167,957,504	1281.69
OPT 350m (QLoRa)	SparseGPT	2:4	0.2395	0.8314	0.2994	0.8467	2271.95	167,957,504	1281.69

Conclusion

- Structured N:M pruning is not structured pruning & worse than unstructured for performance
- Quantization has a drastic effect on runtime
- Activation-based not always better than magnitude-based
- Small models are more sensitive to structured pruning
- Models can hack ROUGE and BertScore metrics by repeating text
- Small decoder-only models struggle to train on tasks with long prompts but short expected outputs, because they end up over-fitting to the prompt.
- QLoRa may not be sufficient to fine-tune small decoder-only models for the summarization task after they have been pruned.