
Sparse Fine-Tuning for Inference Acceleration of Large Language Models

Eldar Kurtic*
IST Austria

Denis Kuznedelev*
Skoltech & Yandex

Elias Frantar
IST Austria

Michael Goin
Neural Magic

Dan Alistarh
IST Austria & Neural Magic

Abstract

We consider the problem of accurate *sparse fine-tuning* of large language models (LLMs), that is, fine-tuning pretrained LLMs on specialized tasks, while inducing sparsity in their weights. We observe that standard loss-based fine-tuning may fail to recover accuracy, especially at high sparsities. To address this, we perform a detailed study of distillation-type losses, determining an L2-based distillation approach we term SquareHead which enables accurate recovery even at higher sparsities, across all model types. On the efficiency side, we show that sparse LLMs can be executed with speedups by taking advantage of sparsity, for both CPU and GPU runtimes. While the standard approach is to leverage sparsity for computational reduction, we observe that in the case of memory-bound LLMs sparsity can also be leveraged for reducing memory bandwidth. We exhibit end-to-end results showing speedups due to sparsity, while recovering accuracy, on T5 (language translation), Whisper (speech translation), and open GPT-type (MPT for text generation). For MPT text generation, we show for the first time that sparse fine-tuning can reach 75% sparsity without accuracy drops, provide notable end-to-end speedups for both CPU and GPU inference, and highlight that sparsity is also compatible with quantization approaches. Models and software for reproducing our results are provided in Section A.

1 Introduction

Large Transformer models [1] have gained high popularity and adoption due to their breakthrough performance across a wide range of challenging tasks. To address their high runtime costs, several acceleration techniques have been developed, e.g., [2, 3, 4, 5]. For inference acceleration, the most popular technique is *quantization*, e.g., [6, 7, 3, 5]: LLMs can be quantized down to 4 bits per weight (or more) almost without loss, and that this can be leveraged for inference speedups, e.g [3, 5]. Yet, quantization methods are reaching accuracy limits at around 3 bits per weight: at this point, accuracy appears hard to recover with current techniques [8, 9, 10].

A key alternative to quantization is *weight sparsity* [11], which consists of setting individual LLM connections to zero. For smaller models, e.g., BERT [12], it is known [13, 14] that high levels of sparsity can be applied during *fine-tuning*, i.e. the process by which the pretrained model is adapted to a “downstream” task, such as question answering or text classification, leading to significant speedups. However, it is not known whether similar techniques work at the scale of LLMs.

In this preliminary work, we study efficient sparse fine-tuning for LLMs, across three modern applications: speech transcription using Whisper [15], specialized to a specific language, machine translation using T5 [16], specialized to a specific language pair [17], and higher-level reasoning using the open GPT-type MPT model [18], specialized on the Grade-School Math (GSM) task [19].

In this context, our contributions are as follows:

*These authors contributed equally.

- We observe that naive sparse fine-tuning [13], which follows dense fine-tuning while gradually imposing sparsity, is challenging to apply for LLMs due to training instability, which manifests itself in several forms: 1) *loss spikes* at higher sparsity values, leading to divergence, 2) *poor recovery*, as the relatively small amount of fine-tuning data available for the subtask may not be sufficient to recover accuracy; or 3) *overfitting*, since iterating multiple times over the limited fine-tuning data leads to low training loss, but high validation error.
- To address this, we investigate fine-tuning sparse models obtained via SparseGPT [20] using various losses which incorporate standard cross-entropy, output knowledge distillation [21], and a type of per-token ℓ_2 knowledge distillation inspired by [22, 23, 24] which we call *SquareHead*. We show that SquareHead distillation consistently recovers accuracy, even at high sparsities.
- On the practical side, we show that the resulting sparse models can be executed with inference speedups. For CPU inference, we leverage the DeepSparse inference runtime [25] to obtain speedups across all three applications. In addition, focusing on generative inference, we present GPU speedups via a GPU-aware N:M sparse format, and remarkable CPU speedups, e.g. of 7.5x with minimal accuracy loss vs the FP32 baseline, jointly leveraging sparsity and quantization.

Together, our results show that with the right set of techniques, sparsity can be successfully applied in the challenging LLM fine-tuning scenario, and that this can lead to significant speedups across both, CPUs and GPUs.

2 Methodology

2.1 Sparse Fine-tuning

Sparsification. To obtain a set of sparse models satisfying target compression requirements, we increase the sparsity level gradually while fine-tuning the model on the task of interest. Unless otherwise stated, we start from a list of desired sparsity levels, in increasing order, and iteratively prune and fine-tune the model while following the original fine-tuning recipe in each subsequent cycle.

Distillation strategies. LLMs are notoriously difficult to train and fine-tune, and we found this to be particularly the case when sparsity is imposed during an often short fine-tuning cycle. Choosing the “right” loss function for fine-tuning is particularly important, and thus we investigate knowledge distillation (KD) approaches. The sparse student model is trained to mimic the behavior of a dense teacher, which has already been fine-tuned on the target task. The most common KD strategy adds a loss term measuring the KL divergence between student and teacher outputs [21]. However, we obtain better results by going further and distilling intermediate representations.

Standard output distillation uses the KL-divergence between student and teacher logits as loss:

$$\mathcal{L}_{\text{logit}} = D_{\text{KL}}(\theta_t || \theta_s) = \frac{1}{\sum_i \mathbb{1}[i \notin \mathbf{P}]} \sum_i^{B \times \text{seq}} \mathbb{1}[i \notin \mathbf{P}] p_{\theta_t}(\mathbf{x}_i) \log \frac{p_{\theta_t}(\mathbf{x}_i)}{p_{\theta_s}(\mathbf{x}_i)}. \quad (1)$$

Above, B stands for batch size, seq for sequence length, $p_{\theta_t}(\mathbf{x}_i)$ and $p_{\theta_s}(\mathbf{x}_i)$ denote output probabilities for teacher and student model respectively. The notation $\mathbb{1}[i \notin \mathbf{P}]$ means that the loss for padding tokens \mathbf{P} is discarded.

To transfer intermediate representations, we examine normalized mean squared error (MSE) loss on each feature representation, which we for simplicity call **SquareHead**:

$$\mathcal{L}_{\text{feat}}^l = \frac{\text{MSE}(f_t^l, f_s^l)}{\text{MSE}(f_t^l, 0)}, \quad \text{shape}(f_s^l) = \text{shape}(f_t^l) = B \times \text{seq} \times d_{\text{model}}. \quad (2)$$

Here, f_t^l and f_s^l denote the feature map of l -th layer of teacher and student model, and MSE represents the mean squared error calculated as $\text{MSE}(X, Y) = \frac{1}{N} \sum_{i=0}^N (x_i - y_i)^2$, for N -dimensional vectors X and Y . We discard terms that correspond to padding tokens. The motivation for the normalization is that the magnitude of activations may greatly vary between different layers, thus focusing the optimization procedure to optimize the layers with largest norm. We observed that for some models MSE loss without normalization leads to instabilities in training. The total feature loss is the sum of per-layer losses over all encoder/decoder blocks.

SquareHead Distillation. The overall SquareHead distillation loss is the sum of original task loss and SquareHead term with equal weights: $\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{feat}}$.

Prior Work. As stated previously, variants of this loss have been used by [22, 23, 24] in the context of compressing smaller-scale models during the fine-tuning process, such as BERT-style models [12] for question-answering or sentiment classification fine-tuning tasks. In the context of BERT fine-tuning, it is known that variants of knowledge distillation can help reduce accuracy loss during fine-tuning, e.g. [13, 14, 24, 23, 26]. Relative to this line of work, in this paper we determine a type of distillation loss that is consistently effective for accurate sparse fine-tuning of large models.

2.2 Runtime Acceleration from Sparsity

We now turn our attention to the question of obtaining practical runtime speedups from a sparse fine-tuned model. The standard strategy is to leverage sparsity *reducing computational load*: since we can skip multiplications with zero, higher sparsity means lower inference computation. However, leveraging the relatively-low sparsities arising in accurate deep models for computational gains is known to be difficult, especially on GPUs [27, 28].

In the case of LLMs, we have a complementary path to sparse acceleration, since these models are often *memory bound* [3, 5]. That is, a significant fraction of the generation runtime can be taken by loading the weights (from memory to registers) to perform layer computation.

To leverage this, we can store sparse weights in compressed form, and decompress them on-the-fly while performing the layer computation. This may even involve multiplication with zeros, as long as they do not have to materialize weights in slow memory. We present results on CPU and GPU runtimes which leverage variants of this approach. For CPU execution, we leverage the DeepSparse runtime [25] which implements sparsity-aware inference in both memory-bound and compute-bound scenarios.

We show speedup results vs. the respective dense baseline, for this 4096×12288 matrix both on CPU and GPU in Figure 1, observing that both implementations achieve substantial speedups. For reference, we can prune the generative MPT-7B model both in 16:32 and 16:64 formats, and fully recover accuracy on GSM8K (see Table 4 in Section 4).

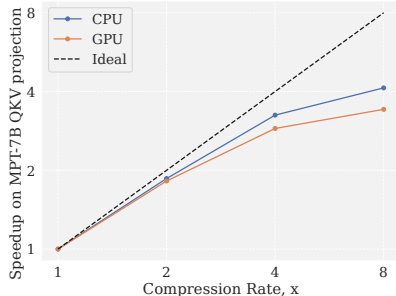


Figure 1: Speedups for sparse CPU and GPU kernels on an MPT-7B layer.

3 Application 1: Compression of T5 and Whisper Translation Models

Experimental Setup. We consider sparsity levels corresponding to 2x, 3x, 4x, 5x, 6x, 7x, 8x, and 10x compression ratios, applied uniformly per layer with the SparseGPT [20] pruner. We compare sparse fine-tuning with three variants of losses. Cross Entropy (original loss, $\mathcal{L}_{\text{task}}$), Standard KD ($\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{logit}}$), and SquareHead KD ($\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{feat}}$). We report hyperparameters in Appendix C. Speedups are reported for end-to-end execution in the DeepSparse inference engine [25] on an Intel Sapphire Rapids CPU with 8 cores (AWS m7i.4xlarge). In the following section, the comparison between distillation strategies (left figure) is done *without* pruning the final output head, whereas the language model head is pruned in the runtime experiments (right tables), since this has non-negligible impact on total inference time.

Compression for language translation using T5. We begin by investigating sparsification of a pretrained T5 model [29], fine-tuned on the popular English-German subset of WMT14 [30]. Following standard practice, we compute BLEU scores on a validation set of the corresponding split as a measure of model’s accuracy. Figure 2 shows accuracy-vs-sparsity trade-off for various loss functions. Table 1 shows scores for the best-performing loss variant (SquareHead KD), together with self-speedups. The dense baseline needs 370ms to encode/decode 128 tokens. We observed that sparse training with Cross Entropy only and Standard KD becomes unstable at high sparsity and only SquareHead KD manages to produce highly-sparse models with reasonable performance. An example of training loss curve is presented in Appendix E.

Compression of speech-to-text using Whisper. In Figure 3 and Table 2, we study compression of Whisper-Small (244M) [31] for Automatic Speech Recognition (ASR) on the Hindi (Hi) language

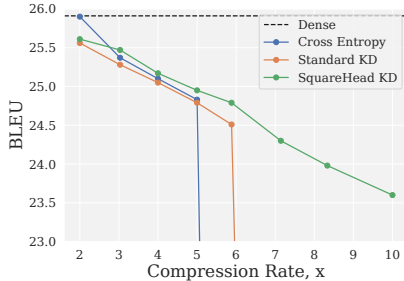


Figure 2: BLEU (\uparrow) score for variants of loss functions on English-German WMT14 and T5-Small at various compression rates.

Model	Sparsity (%)	BLEU (\uparrow)	Speedup (\times)
T5	0	25.91	1.00
	50	25.44	1.89
	67	25.04	2.12
	75	24.67	2.14
	80	24.57	2.16
	83	24.31	2.22
	86	24.11	2.25
	88	23.47	2.38
	90	23.10	2.42

Table 1: BLEU score and speedups of T5-Small for various sparsities and SquareHead KD loss on English-German WMT14.

subset of CommonVoice 11.0 [32]. We report Word Error Rate (WER) as a standard measure of ASR performance. Our dense CPU baseline takes 882ms to transcribe an audio sequence of 15 seconds and a detailed breakdown is presented in Appendix F.

Discussion. Results show that the SquareHead KD loss improves upon standard approaches in these two fine-tuning scenarios, especially at higher sparsities, where the latter tend to diverge. (We provide an analysis of this phenomenon in Appendix D.) Overall, we can induce moderate sparsity levels (50-70%) while preserving or even improving accuracy, leading to speedups of 1.5-2x. We can obtain higher speedups, of up to 2.67x, with 80-90% sparsity, at the price of higher error rates.

4 Application 2: Compression of Generative Models

Next, we investigate compression of generative models, specifically the open-source Mosaic pre-trained model MPT-7B [18]. We focus on GSM8K [33], a dataset with high quality and diverse grade school math problems, following the recipe of [34]. On this task, in zero-shot mode, the baseline model completely fails, with a score of 0%, whereas in 8-shot evaluation it scores only 6.8%. These results suggest that the model necessitates additional refinement via supervised fine-tuning (SFT).

Experimental setup. First, we fine-tune MPT-7B via SFT to obtain a highly accurate and competitive dense baseline, which we use as the teacher in knowledge distillation runs. Then, we apply oneshot unstructured pruning with SparseGPT to 40%, 50%, 60%, 70%, and 80% sparsity targets, uniformly across all layers, which correspond to 1.7x, 2.0x, 2.5x, 3.3x, and 5.0x compression ratios respectively. We explore how different sparse fine-tuning techniques help in recovering accuracy of the dense baseline model. We provide detailed description of hyper-parameters in Appendix C.2. After fine-tuning, we investigate compatibility of sparse models with quantization via post-training quantization to INT8. To achieve this we leverage the SparseML [35] library and quantize to 8-bits the weights and activations of all linear weight matrices, and two batch matrix multiplications in attention layers. For accuracy evaluation on the GSM8K task we utilize the standardized evaluation protocol via Language Model Evaluation Harness [36].

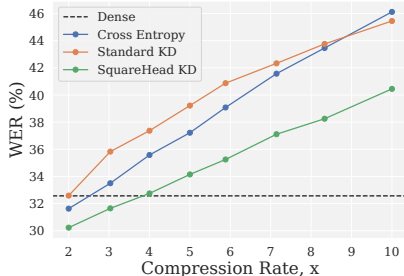


Figure 3: WER (\downarrow) for variants of loss functions on Hindi dataset and Whisper-Small at various compression rates.

Model	Sparsity (%)	WER (\downarrow)	Speedup (\times)
Whisper	0	32.57	1.00
	50	30.87	1.58
	67	31.85	2.15
	75	33.14	2.23
	80	34.69	2.35
	83	35.77	2.49
	86	37.33	2.52
	88	38.64	2.58
	90	40.60	2.67

Table 2: WER and speedups of Whisper-Small for various sparsities and SquareHead KD loss on Hindi dataset.

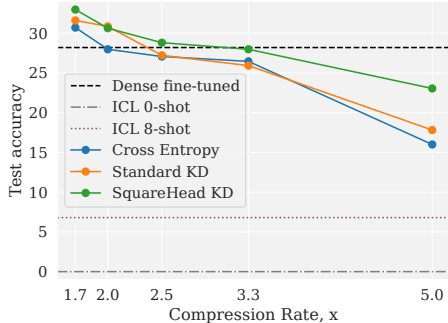


Figure 4: Test accuracy for variants of loss functions on GSM8K dataset and MPT-7B model at various compression rates. ICL stands for in-context learning results using the pretrained dense model, without fine-tuning.

N:M format	Sparsity (%)	Test accuracy
dense	0	28.2
2:4	50	31.4
16:32	50	31.4
16:64	75	28.6

Table 4: Test accuracy of pruned MPT-7B models on GPUs via N:M sparsity format.

Machine	Num Cores	Dense	60%	70%	80%	60%	70%	80%
		FP32	FP32	FP32	FP32	INT8	INT8	INT8
Xeon Gold 6430	1	0.6	1.2	1.5	2.1	3.9	4.7	6.1
	4	2.1	4.3	5.5	8.2	14.1	16.3	19.6
Ryzen 9 7950X	1	1.4	2.0	2.0	2.1	7.2	7.7	7.9
	4	2.5	5.4	6.3	7.6	17.4	20.9	26.7

Table 5: Latency performance (tokens/second) for the MPT-7B model at various sparsities, using DeepSparse [25] on 1 and 4 cores, using Intel and AMD CPUs.

Discussion. The accuracy results for different losses are shown in Figure 4. They exhibit very similar trends to the prior two applications, showing that SquareHead KD is superior to both standard CE loss and standard KD for sparse fine-tuning. The fact that the distilled models can outperform the dense baseline at low sparsity can be explained due to the effect of distillation being applied, but also possibly the regularizing effect of low sparsity. With SquareHead KD, we can obtain FP32 models with 70% and 75% sparsities which have essentially no loss (in terms of test accuracy) relative to the dense model, even though pruning is performed in one shot. We emphasize that fine-tuning with SquareHead KD improves the dense model’s accuracy as well, from 28.2 to 33.0.

We now examine the speedup-vs-accuracy trade-offs, presented in Table 3 for both FP32 and INT8 models. In FP32, moderate sparsities (60-70%) can be reached losslessly, leading to speedups of 2-2.5x. Moving to INT8, we observe a consistent accuracy decrease of 1-2% at each sparsity level, due to post-training quantization. At the same time, this loss of accuracy is accompanied by a major performance improvement, since the gains due to these two compression approaches compound. In absolute terms, the 70% INT8 model can execute at a remarkable **7.7 tokens/second on a single core** of an AMD Ryzen CPU, and at **20.9 tokens/second** on 4 cores. The speedup of the lossless 60%-sparse INT8 model is of approximately 6.7x, and we can reach a decoding speedup of 9.08x at 80% INT8, at the price of a 7% drop in accuracy.

5 Discussion

We have shown early results suggesting that sparsity can be an effective acceleration approach for LLM inference, focusing on “modern” applications such as speech and language translation, and text generation following complex instructions. Importantly, our study shows that sparsity is compatible with quantization in the memory-bound generative setting, and that together these techniques can lead to remarkable performance for computationally-limited devices. Based on prior work, we have identified a general distillation approach to recover accuracy while inducing high sparsity over limited fine-tuning data. In future work, we plan to expand this preliminary study by exploring sparse fine-tuning for larger-scale models and tasks in the generative setting, higher quantization degrees, as well as the practically-relevant setting of pruning on the pretraining task [37], followed by fine-tuning the already-sparsified model on a specialized dataset.

Sparsity (%)	FP32		INT8	
	Test accuracy	CPU speedup	Test accuracy	CPU speedup
0	28.2	1.00x	27.8	3.98x
40	32.9	1.54x	30.3	5.31x
50	30.6	1.78x	30.7	5.72x
60	28.8	2.07x	28.4	6.70x
70	28.0	2.62x	27.1	7.49x
80	23.1	3.35x	21.1	9.08x

Table 3: Test accuracy of pruned (FP32) and pruned-quantized (INT8) MPT-7B models on the GSM8K dataset. Speedups are measured relative to FP32, for end-to-end decode latency in DeepSparse [25] using an 8 core CPU (AWS m7i.4xlarge) at sequence length 512.

Acknowledgments

The authors thank the Neural Magic Machine Learning Research team for useful discussions during the development of this paper. We would also like to thank Eugenia Iofinova for useful comments on an earlier version of this draft, and Artur Niederfahrenheit for useful suggestions regarding fine-tuning on the GSM dataset.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.
- [2] Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with io-awareness. *arXiv preprint arXiv:2205.14135*, 2022.
- [3] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.
- [4] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*, 2022.
- [5] Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. *arXiv preprint arXiv:2212.09720*, 2022.
- [6] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.
- [7] Guangxuan Xiao, Ji Lin, Mickael Seznec, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. *arXiv preprint arXiv:2211.10438*, 2022.
- [8] Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *arXiv preprint arXiv:2206.01861*, 2022.
- [9] Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023.
- [10] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. Quip: 2-bit quantization of large language models with guarantees. *arXiv preprint arXiv:2307.13304*, 2023.
- [11] Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Conference on Neural Information Processing Systems (NeurIPS)*, 1990.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.
- [13] Victor Sanh, Thomas Wolf, and Alexander M. Rush. Movement pruning: Adaptive sparsity by fine-tuning. *arXiv preprint arXiv:2005.07683*, 2020.
- [14] Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, and Dan Alistarh. The Optimal BERT Surgeon: Scalable and accurate second-order pruning for large language models. *arXiv preprint arXiv:2203.07259*, 2022.
- [15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.
- [16] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- [17] Matouš Macháček and Ondřej Bojar. Results of the wmt14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, 2014.

- [18] MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms. www.mosaicml.com/blog/mpt-7b, 2023. Accessed: 2023-10-01.
- [19] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [20] Elias Frantar and Dan Alistarh. Massive language models can be accurately pruned in one-shot. *arXiv preprint arXiv:2301.00774*, 2023.
- [21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [22] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.
- [23] Eldar Kurtic, Elias Frantar, and Dan Alistarh. Ziplm: Hardware-aware structured pruning of language models. *arXiv preprint arXiv:2302.04089*, 2023.
- [24] Elias Frantar and Dan Alistarh. SPDY: Accurate pruning with speedup guarantees. *arXiv preprint arXiv:2201.13096*, 2022.
- [25] NeuralMagic. DeepSparse. <https://github.com/neuralmagic/deepsparse>, 2021.
- [26] Mengzhou Xia, Zexuan Zhong, and Danqi Chen. Structured pruning learns compact and accurate models. *arXiv preprint arXiv:2204.00408*, 2022.
- [27] Trevor Gale, Matei Zaharia, Cliff Young, and Erich Elsen. Sparse gpu kernels for deep learning. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14. IEEE, 2020.
- [28] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *arXiv preprint arXiv:2102.00554*, 2021.
- [29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.
- [30] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.
- [31] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.
- [32] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.
- [33] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [34] Artur Niederfahrenheit, Kourosh Hakhamaneshi, and Rehaan Ahmad. Fine-Tuning LLMs: LoRA or Full-Parameter?, 2021.
- [35] Mark Kurtz, Justin Kopinsky, Rati Gelashvili, Alexander Matveev, John Carr, Michael Goin, William Leiserson, Sage Moore, Bill Nell, Nir Shavit, and Dan Alistarh. Inducing and exploiting activation sparsity for fast inference on deep neural networks. In *International Conference on Machine Learning (ICML)*, 2020.
- [36] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation. <https://github.com/EleutherAI/lm-evaluation-harness>, 2021.

[37] Elias Frantar, Carlos Riquelme, Neil Houlsby, Dan Alistarh, and Utku Evci. Scaling laws for sparsely-connected foundation models. *arXiv preprint arXiv:2309.08520*, 2023.

A Reproducibility

To promote reproducibility of our results, we provide the following resources:

- Code for sparse fine-tuning of T5 and Whisper models can be found here: <https://github.com/IST-DASLab/TACO4NLP>.
- Code for sparse fine-tuning of GPT-type (e.g. MPT) models can be found here: <https://github.com/IST-DASLab/SparseFinetuning>.
- MPT models are released at <https://sparsezoo.neuralmagic.com/?datasets=gsm8k&ungrouped=true>
- CPU speedup for generative inference can be reproduced by following the instructions at <https://github.com/neuralmagic/deepsparse/tree/main/research/mpt> and as a HuggingFace Collection.
- GPU code will be provided at <https://github.com/IST-DASLab/sparsegpt/>.

B GPU Implementation Sketch

For GPU inference, we show speedup potential via the following implementation for “uniform” sparsity patterns such as 16:32 (50%), 16:64 (75%) and 16:128 (87.5%). (For clarity, 16:32 means 16 non-zeros in every block of 32 consecutive weights.) We implement a custom CUDA kernel where each threadblock fetches tiles of $16 \times \text{\#threads}$ weights to shared memory and then accumulates partial matrix-vector product outputs by unpacking the corresponding INT32 sparsity bitmasks to determine when weights should be multiplied with inputs.

Executed on an NVIDIA A6000 GPU for a 4096×12288 matrix (shapes of the QKV projection layer in MPT-7B model), our kernel yields $1.82\times$ speedup over dense FP16 execution, which is close to the theoretical expectation of $1.77\times$ (equivalent to 9 bits per weight) for a bitmask representation of sparsity.

C Sparse Fine-tuning Hyperparameters

C.1 Parameters for Whisper and T5 Experiments

In Table 6 we report hyperparameters for sparse fine-tuning of Whisper and T5 models.

Table 6: Hyperparameters used in the sparse fine-tuning experiments with Whisper and T5 models.

Model	Learning rate	Num epochs	Warmup steps	Batch size	Schedule	Weight decay
T5-Small	$2 \cdot 10^{-3}$	3	300	128	linear	10^{-4}
Whisper-Small	$2 \cdot 10^{-4}$	6	50	32	linear	10^{-4}

C.2 Parameters for MPT Experiments

For MPT-7B model, we untie input embeddings and language modelling head for compatibility with quantization via SparseML, where we quantize weights of input embeddings, and weights and activations of the language modelling head. For all sparsities, we oneshot prune the model with the default SparseGPT [20] parameters and then fine-tune for either 2 (40%, 50%, and 60% sparsity) or 4 epochs (70% and 80% sparsity). We use linearly decaying learning rate (LR) with warmup of 20 steps, and sweep for all loss variants over the following peak-LR values: $3e-5$,

5e-5, 8e-5, 1e-4. We use batch-size of 32, and equal weight $\lambda = 1.0$ for losses in compound loss variants. For more details on reproducibility, please check our GitHub repository <https://github.com/IST-DASLab/SparseFinetuning>.

D Entropy Analysis

A possible explanation for the success of SquareHead could be the fact, that fine-tuning with task loss only, especially on limited data, leads to overfitting. We measured the entropy of the predictive distribution of sparse Whisper models and observed that models fine-tuned without knowledge distillation have very low entropy, i.e make overconfident predictions, whereas SquareHead induces regularization.

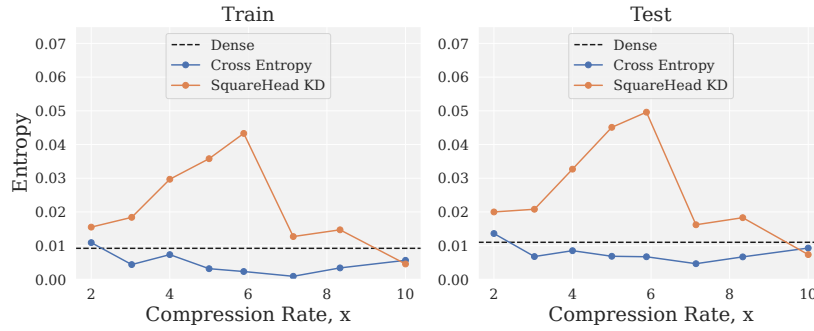


Figure 5: Entropy of the predictive distribution of sparse Whisper-244M models for different fine-tuning approaches on train and test split of Hindi subset of Common Voice dataset.

E Divergence in training T5 on WMT14 dataset

We observed that sparse fine-tuning of T5 becomes unstable at high sparsity when training with task loss only or standard KD. Training loss after pruning step decreases for some number of steps and then suddenly diverges as shown on Figure 6. This issue is encountered even when trying different random seeds. Behavior is similar for training with CE and standard KD. Only training with SquareHead loss appears to be stable at high sparsity.

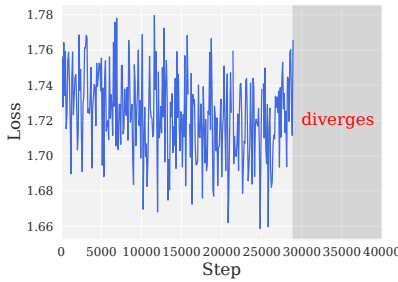


Figure 6: Loss history for diverged T5 run.

F T5 and Whisper Encoder-Decoder Performance Breakdown

Understanding the individual performance of the encoder and decoder components in sequence-to-sequence models like T5 and Whisper is paramount. These components handle different tasks; the encoder processes input data into a context-rich representation, and the decoder translates this representation into generated tokens, using both the encoder’s representation and it’s own context to produce subsequent output. While individual component performance provides insight

into bottlenecks, it’s also essential to evaluate end-to-end performance as it offers a more holistic understanding of how compression affects real-world application scenarios. In Figures 7 and 8, and Tables 7 and 8 we present detailed performance breakdown from the DeepSparse engine.

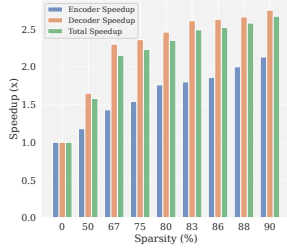


Figure 7: Component speedups of Whisper-Small for various sparsities.

Sparsity (%)	Encoder		Decoder		Total	
	Speedup	Latency	Speedup	Latency	Speedup	Latency
0	1.00x	89.39 ms	1.00x	12.82 ms	1.00x	882.70 ms
50	1.18x	75.49 ms	1.65x	7.76 ms	1.58x	558.21 ms
67	1.43x	62.50 ms	2.30x	5.57 ms	2.15x	410.06 ms
75	1.54x	58.03 ms	2.36x	5.43 ms	2.23x	395.53 ms
80	1.76x	50.87 ms	2.46x	5.21 ms	2.35x	374.93 ms
83	1.80x	49.77 ms	2.61x	4.91 ms	2.49x	354.59 ms
86	1.86x	48.01 ms	2.63x	4.87 ms	2.52x	350.16 ms
88	2.00x	44.66 ms	2.66x	4.81 ms	2.58x	342.70 ms
90	2.13x	41.97 ms	2.75x	4.66 ms	2.67x	330.35 ms

Table 7: Encoder, decoder, and total timings of Whisper-Small for various sparsities. The total workload is generating 60 tokens from 15 seconds of audio.

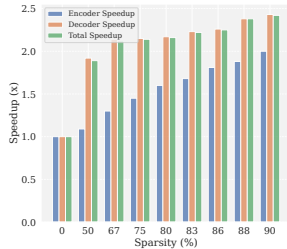


Figure 8: Component speedups of T5-Small for various sparsities.

Sparsity (%)	Encoder		Decoder		Total	
	Speedup	Latency	Speedup	Latency	Speedup	Latency
0	1.00x	5.18 ms	1.00x	2.99 ms	1.00x	391.92 ms
50	1.09x	4.75 ms	1.92x	1.56 ms	1.89x	206.93 ms
67	1.30x	3.97 ms	2.14x	1.40 ms	2.12x	184.80 ms
75	1.45x	3.58 ms	2.15x	1.39 ms	2.14x	183.51 ms
80	1.60x	3.23 ms	2.17x	1.38 ms	2.16x	181.18 ms
83	1.68x	3.08 ms	2.23x	1.34 ms	2.22x	176.32 ms
86	1.81x	2.86 ms	2.26x	1.33 ms	2.25x	174.21 ms
88	1.88x	2.75 ms	2.38x	1.26 ms	2.38x	164.96 ms
90	2.00x	2.59 ms	2.43x	1.23 ms	2.42x	161.84 ms

Table 8: Encoder, decoder, and total timings of T5-Small for various sparsities. The total workload is encoding 128 tokens and generating 128 tokens.

G T5 inference examples

Below in Table 9 we present a couple of non cherry-picked examples of English-German translation with dense and sparse T5 models. Sparse model is pruned to 75%. One can see, that sparse models yields outputs of superficially the same quality at the base dense model.

Table 9: Examples of T5 English-German translation.

Prompt	Dense model	Sparse model
The quick brown fox jumps over the lazy dog.	Der braune Fuchs springt über den faulen Hund.	Der schnelle braune Fuchs springt über den faulen Hund.
I don't respect anybody who can't tell the difference between Pepsi and Coke.	Ich respektiere niemanden, der den Unterschied zwischen Pepsi und Coke nicht erzählen kann.	Ich respektiere niemanden, der den Unterschied zwischen Pepsi und Coke nicht erzählen kann.
Sometimes it is better to just walk away from things and go back to them later when you're in a better frame of mind.	Manchmal ist es besser, einfach weg von den Dingen zu gehen und zu ihnen später zurück, wenn Sie in einem besseren Geistesrahmen sind.	Manchmal ist es besser, sich von den Dingen zu entfernen und später zu ihnen zurückzukehren, wenn Sie in einem besseren Geistesrahmen sind.