Uncovering the Hidden Cost of Model Compression

Diganta Misra *
Carnegie Mellon University, Landskape AI
digantam@andrew.cmu.edu

Muawiz Chaudhary *
Mila - Quebec AI Institute, Concordia University

Agam Goyal *
University of Wisconsin-Madison

Bharat Runwal *
Mila - Quebec AI Institute

Pin Yu Chen
IBM Research

Abstract

In an age dominated by resource-intensive foundation models, the ability to efficiently adapt to downstream tasks is crucial. Visual Prompting (VP), drawing inspiration from the prompting techniques employed in Large Language Models (LLMs), has emerged as a pivotal method for transfer learning in the realm of computer vision. As the importance of efficiency continues to rise, research into model compression has become indispensable in alleviating the computational burdens associated with training and deploying over-parameterized neural networks. A primary objective in model compression is to develop sparse and/or quantized models capable of matching or even surpassing the performance of their over-parameterized, full-precision counterparts. Although previous studies have explored the effects of model compression on transfer learning, its impact on visual prompting-based transfer remains unclear. This study aims to bridge this gap, shedding light on the fact that ***model compression detrimentally impacts the performance of visual prompting-based transfer****, particularly evident in scenarios with low data volume. Furthermore, our findings underscore the adverse influence of sparsity on the calibration of downstream visual-prompted models. However, intriguingly, we also illustrate that such negative effects on calibration are not present when models are compressed via quantization. This empirical investigation underscores the need for a nuanced understanding beyond mere accuracy in sparse and quantized settings, thereby paving the way for further exploration in Visual Prompting techniques tailored for sparse and quantized models.*

1. Introduction

The evolution of deep learning has shifted from training task-specific models to extensive task-agnostic pre-training. The objective is to construct a model with robust universal representations, facilitating numerous downstream tasks without necessitating intensive training. This category of models is now encompassed by the term "Foundation models" (FM) [6]. The efficacy of these pre-trained models for downstream tasks has often been exemplified through straightforward and computationally efficient adaptation methods, such as linear probing (LP) [1], within the broader framework of transfer learning. Transfer learning has traditionally been fundamental for adapting pre-trained models to various downstream tasks, historically constrained to full fine-tuning (FF) and LP. The former, while generally superior in performance, is also the more costly approach, whereas the latter is a more computationally cheaper option but typically exhibits lower performance compared to full fine-tuning methods.

Although linear probing (LP) and full fine-tuning (FF) have traditionally served as standard transfer learning methods, a novel approach, referred to as model reprogramming [8, 15, 58] or more commonly known as visual prompting (VP) [4, 62, 69], has emerged as a viable and efficient alternative, capable of rivaling LP in both performance and transfer cost. Essentially, VP involves learning a perturbation such that, when applied to input samples of the target dataset, the pre-trained model accurately classifies the resulting "reprogrammed" sample without requiring any changes to the weights. Reprogramming relies on aligning the features of the target data with those of the source data, eliminating the need for gradient updates to the network weights. Consequently, VP often competes with LP in terms of efficiency.

While several studies in the literature have empirically evaluated the performance of Linear Probing (LP) and Visual Prompting (VP) across diverse target data scenarios [4, 8, 63, 65, 67], there has been limited exploration of the distinctions between the two methods when subjected to the constraints of (a) *low data volume* and (b) *model compression*. This study aims to reveal the *hidden costs* associated with compressed models in the domain of transfer learning across various data volume settings. In the existing literature,

*equal contribution

Figure 1. When examining the label mapping [7] of the ResNet-50 Sparse LT model [19, 33] alongside its dense counterpart across target classes within the OxfordPets [55] and DTD [10] datasets, a notable distinction emerges: the dense model exhibits a more semantically accurate label mapping. In contrast, the sparse variant often assigns target classes to unrelated classes from the source dataset. This trend echoes similarly in the context of quantization, where the full-precision DeiT (32-bit) [60] demonstrates superior semantic accuracy and consistency in label mapping compared to its quantized counterpart (2-bit) [66] across various target classes within the OxfordPets and DTD datasets.

these costs span a broad spectrum, including model attributes such as reliability [5, 26], performance under distribution shifts[9, 45], fairness [37, 38, 61], and more.

To this end, under the constraint of low data volume, we examine the transferability of pre-trained models in few-shot settings. For the constraint of model compression, our investigation encompasses a wide array of sparse and quantized models generated through compression techniques such as unstructured pruning [31, 32, 43], structural pruning [50, 70], quantization [12, 13, 20, 66, 68], and solutions based on the Lottery Ticket Hypothesis (LTH) [19].

Moreover, we expand our empirical framework to investigate the influence of transferring compressed models using VP on the calibration of the resulting model in the target task, comparing it to its dense counterparts. Although recent research has prioritized the comprehension and enhancement of fairness and reliability in models [69], along with the examination of pruning strategies and the resultant models [2, 25, 44, 47, 64], there remains a gap in our understanding of the reliability of visual prompting as an adaptation mechanism, particularly in the context of varying model compression rates.

Our main contributions can be summarized as follows.

1. We conduct a comprehensive empirical investigation into the effects of model compression and low data volume on the transferability of pre-trained models through visual prompting methods.
2. Our in-depth analysis across eight datasets reveals a notable decrease in performance for compressed models compared to their dense counterparts when transferred via visual prompting.
3. Significantly, we empirically examine the adverse impact of employing sparse models on the calibration of the final model obtained after transfer to the downstream target task using visual prompting, marking the first exploration of this phenomenon.
4. Moreover, in contrast, we illustrate that quantized models do not suffer from the adverse effects on calibration observed in sparse models. We offer an intuitive analysis to elucidate this phenomenon.
5. To that extent, we provide a fine-grained overview of our observations based on the categorization of different compression methods in Table 1.
6. Furthermore, we contribute novel insights into the distinctions between visual prompts and their corresponding label mapping learned by compressed models compared to their dense, full-precision counterparts.

2. Background and Related Works

Visual Prompting and Reprogramming. Model reprogramming [8, 15] is a novel parameter-efficient fine-tuning method that integrates two trainable modules, the input transformation layer and the output mapping layer, into a pre-trained model for transfer learning. During the reprogramming training process, only the parameters linked to the inserted layers are updated, maintaining the unchanged parameters of the pre-trained model. A significant advantage of model reprogramming lies in its proficiency in cross-domain learning, enabling successful application in diverse domains such as biomedical image classification [62], time series classification in speech models [67], and protein sequence learning in language models [65]. Specifically, in image classification tasks within the same domain, model reprogramming equates to visual prompting (VP) [4]. In simple terms, VP achieves input transformation through a universal trainable additive padding operation for each image and output mapping via a function specifying the transition

from source label classes to target label classes. To realize the output mapping, [62] proposed frequency-guided label mapping, while [7] proposed iterative label mapping for VP, known as ILM-VP. More description about VP are provided in supplementary material.

Model Compression and Transfer Learning. Model compression enhances inference efficiency by reducing memory requirements, with pruning and quantization standing out as prominent methods.

Pruning: Achieving model sparsity, as introduced by [31, 43], effectively compresses over-parameterized deep neural networks. Progressive sparsification methods like GMP (Gradual Magnitude Pruning) [23, 32] and sparse regularization techniques such as AC/DC [56] and RigL [16] showcase dynamic approaches to weight pruning. The Lottery Ticket Hypothesis (LTH) [19] identifies sparse subnetworks within large networks, offering comparable or superior performance. Recent work, like Upop [59], addresses pruning limitations for large vision-language models. *Quantization:* A widely used compression technique, quantization [24], involves representing weights or activations in lower precision. It includes quantization-aware training, conducted during fine-tuning or retraining, and post-training quantization applied after model training. Recent studies exploring quantization in LLM aim to reduce operation costs [12, 20, 68], albeit often at the expense of performance. For Vision Transformers, recent work [66] identifies unique variation behaviors in ViT, distinct from CNNs, and proposes an efficient knowledge distillation-based variation-aware quantization method to address this issue.

In transfer learning with sparse models, [40] investigated various pruning techniques' effects on downstream performance using sparse pre-trained ImageNet models, comparing linear probing and fine-tuning for transfer. However, there is limited detailed study for quantized models in this context. Additionally, [21] introduced adversarial robustness into the transfer learning pipeline for improved transferability, while [28] explored the robustness of quantized models.

Calibration of Neural Networks. [29] first identified neural networks' tendency for overconfident predictions, a phenomenon exacerbated in sparse training, such as RigL [16] as noted by [44]. While efforts focus on efficiency and adaptation speed in pruned models [40], sparse model reliability in transfer learning remains unexplored. Given their broad application and transfer across domains [71], overconfidence poses safety risks in critical domains like self-driving and healthcare [35, 36]. Recent studies address calibration in pruned models [2], propose sparse training improvements for lottery tickets [44], and integrate calibration into pruning techniques [64].

Although visual prompt-based transfer has demonstrated promising outcomes across various downstream tasks, the dependability of this innovative method when employed with sparse models remains largely unexplored.

3. Results

Compression	Type	Models	Performance Drop
GMP [30, 49]	Unstructured	ResNet-(18,34)	High
IMP (LT) [19]	Unstructured	ResNet-50	Moderate
AC/DC [56]	Unstructured	ResNet-50	Low
RigL [16]	Unstructured	ResNet-50	Low
UPop [59]	Structured	CLIP ViT-L	High
VVTQ [66]	Quantization	DeiT-T, Swin-T	Moderate

Table 1. **Summary of Results.** Key Observations about the performance drop in various architectures of models using different compression strategies and transferred via Visual Prompting.

In this section, we look at results from our extensive analysis with the aim of understanding the difference between the downstream performance of various visual prompting methods in terms of performance under conditions of both low data volumes and model compression, as well as reliability in terms of calibration under different model compression rates.

3.1. Experimental Setup

We consider eight target datasets that encompass a mixture of downstream tasks in the near- and far domain, namely CIFAR-10 [41], SVHN [53], GTSRB [39], DTD [10], Flowers102 [54], OxfordPets [55], EuroSAT [34] and Caltech101 [46]. In terms of model architectures, we base our experiments on the ResNet [33] family of models with ResNet-18, ResNet-34, ResNet-50 (Sec. 3.2.1), CLIP [57] (Sec. 3.2.2), and Vision Transformer [14] models with DeiT-T and Swin-T [51, 60] (Sec. 3.2.3).

For pruned models, we use a GMP-pruned model [30, 49] for ResNet-18 and ResNet-34, and AC/DC [56], RigL [16], and solutions of the lottery ticket hypothesis (LTH)[1] [19] for ResNet-50 derived at different sparsity levels. All of these checkpoints are pre-trained on the ImageNet-1k [11] classification task. As shown in Fig. **??**, most lottery ticket solutions demonstrate superior performance in terms of accuracy than that of the parent network on the pre-training dataset.

For quantized models, we use a VVTQuantized model [66] for DeiT-T and Swin-T models at varying bits of quantization on both activations and weights, from full 32 bit to 2 bit. All of these checkpoints are pre-trained on the ImageNet-1k [11] classification task. As mentioned in [66], quantized 4 bit DeiT-T and Swin-T models obtain better performance than it's full-precision counterpart.

To ensure consistency and measure statistical significance, all configurations were run with three seeds, amounting to

[1] Solutions of LTH are essentially LT initializations that are trained to convergence.

more than 15,000 experiments in total. Finally, for method of visual prompting (VP), we will be demonstrating our results for VP based on three popular label mapping techniques of Random Label Mapping (RLM-VP), Fequency-based Label Mapping (FLM-VP) [62], and mainly Iterative Label Mapping (ILM-VP) [7] which is the state-of-the-art VP method. Furthermore, for our study on the performance of sparse foundation models (Sec. 3.2.2), we also compared ILM-VP with the implementation of visual prompting in [4]. Finally, we investigate the cross-modal reprogramming [52] for various models and compression settings in Sec. 3.2.4.

Note: We define some terms and notations that will be used throughout this section:

- Δ in the LTH heat-maps (**??**) refers to the difference between the accuracy of the dense model and the corresponding LT at the specified sparsity and data budget. This same notation is also used in our analysis of the class-wise impact on the performance of sparse model transfer (**??**).
- $\mathcal{S}$ or "%-sparsity" in all plots refers to the percentage of remaining weights in the model or, in other words, the capacity of the sparse model.
- "N-shots" refers to the training data budget, i.e. the number of samples per class of the downstream dataset used during training.

3.2. Performance Analysis

3.2.1 Sparse Models

GMP: Analyzing the transfer performance of GMP-pruned [72] ResNet-18 and ResNet-34 models at around a $80-90\%$ layer-wise sparsity level on various downstream datasets using ILM-VP, FLM-VP, and RLM-VP, we can see a clear detrimental impact of sparse models compared to their equivalent dense counterparts.

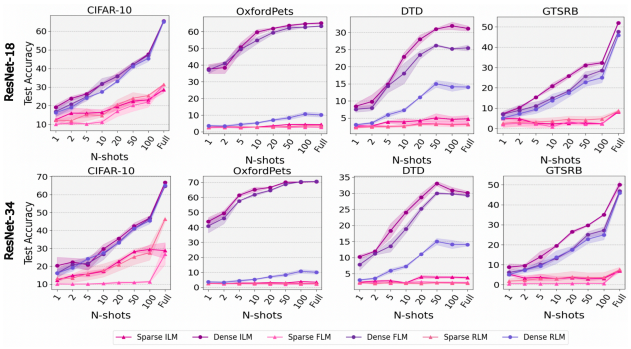

Figure 2. **GMP-pruned ResNet-18/34.** Transfer performance measured by test accuracy of pruned ResNet-18/34 model on a variety of downstream datasets and varying levels of data budgets.

In Figure 2, ResNet-18 analysis shows ILM-VP is generally the most effective for dense models, followed by FLM-VP. RLM-VP consistently lags, except in GTSRB where all VP methods perform similarly. Sparse models across settings consistently underperform dense counterparts, with no clear trend among different VP modes. Pruned model transfer impact is most significant in OxfordPets, with a consistent performance gap exceeding 50% for various data budgets. Similarly, for ResNet-34 in Figure 2, trends echo ResNet-18. Dense models outperform sparse counterparts across all data budgets and downstream datasets, with a notable accuracy improvement, especially in full data settings, perhaps because of the improved size of the model architecture compared to ResNet-18. ILM-VP remains consistently superior for both sparse and dense models.

In summary, we observe the detrimental impact of transfer via visual prompting methods on ResNet-18 and ResNet-34 models pruned using GMP across multiple downstream datasets and varying data budget settings, despite the models matching the dense model's upstream ImageNet-1k performance with 69.8% for ResNet-18 and 73.3% for ResNet-34.

AC/DC and RigL: We now study the transfer performance for pruned ResNet-50 models at $80\%, 90\%$ and 98% sparsity compressed by AC/DC [56] and at $80\%, 90\%$ and 95% sparsity compressed by RigL [16].

In Figures 3 and 4, the dense model generally outperforms sparse models for both AC/DC and RigL across four datasets. An exception is observed for GTSRB, where sparse models in specific sparsity and data budget settings match the performance of their dense counterpart. Comparing with the previously pruned ResNet-18 and 34 models using GMP, the ResNet-50 models pruned with AC/DC and RigL, while still generally worse than their dense counterparts, exhibit relatively better performance. The overall trends in performance on downstream tasks for models pruned by these dynamic sparsification techniques are quite similar.

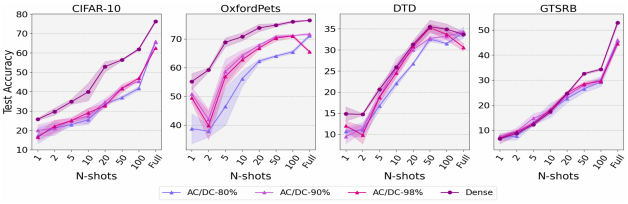

Figure 3. **AC/DC-pruned ResNet-50.** Transfer performance measured by test accuracy of pruned ResNet-50 model on a variety of downstream datasets and varying levels of data budgets.

We also examine the performance of Lottery Ticket Hypothesis (LTH) solutions for ResNet-50 when transferred at various sparsity configurations and data-budget settings in the supplementary material. Our conclusions primarily rely on the trends for ILM-VP, the sota method. In general, it is evident that the transfer of these LTH solutions using VP-based methods does not maintain their performance under low data volumes, despite their upstream performance matching or outperforming their dense counterparts.

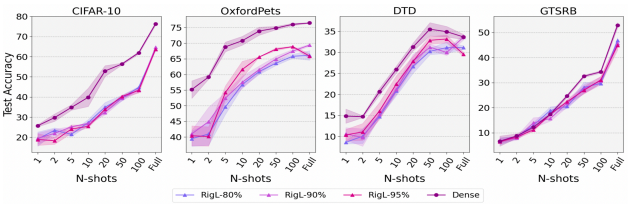

Figure 4. **RigL-pruned ResNet-50.** Transfer performance measured by test accuracy of pruned ResNet-50 model on a variety of downstream datasets and varying levels of data budgets.

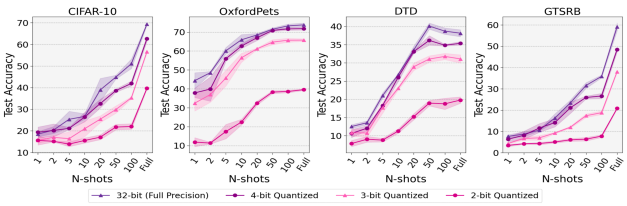

Figure 5. **VVTQuantized DeiT-T.** Transfer performance measured by test accuracy of quantized DeiT-T models on a variety of downstream datasets and varying levels of data budgets.

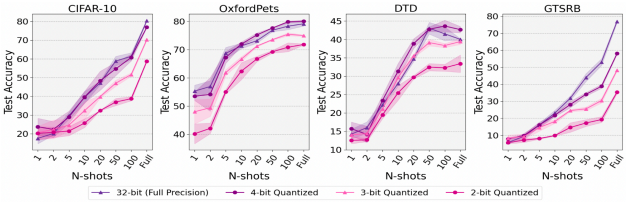

Figure 6. **VVTQuantized Swin-T.** Transfer performance measured by test accuracy of quantized Swin-T models on a variety of downstream datasets and varying levels of data budgets.

3.2.2 Sparse Foundation Models

This section presents the impact of visual prompting on compressed CLIP models. Specifically, we study the effect on transfer performance on compressed CLIP ViT-Large [57], where we use two compressed models: 2x and 4x provided by UPop [59] for the image-retrieval task on the COCO dataset. UPop, a structured pruning framework for vision language transformers, adaptively allocates pruning ratios to selected model components and employs progressive pruning to achieve substantial compression ratios.

We present the results of Visual Prompting (VP) [4] and ILM-VP [7] techniques applied to CLIP across seven datasets (see Table 2). We trained the visual prompt for 10 epochs with a batch size of 16 using an SGD optimizer and a cosine learning rate scheduler. Further details of the setup are available in the supplementary material. The table indicates that the transfer performance of the dense model is optimal compared to the compressed counterparts across all datasets, and the 4x compressed model exhibits the weakest performance. Although this effect is noticed to a lower extent on datasets like EuroSAT and GTSRB where the performance drop is only $\sim$ 1-2%, on datastes like Caltech101 and DTD this gap is much more pronounced with an average drop of 20% and is exacerbated even further for OxfordPets

Datasets	Method	Dense (856.0M)	2x (473.7M)	4x (280.2M)
CIFAR-10	VP	**97.31 %**	92.65 % (4.66 ↓)	90.07 % (7.24 ↓)
	ILM-VP	**97.53 %**	93.04 % (4.49 ↓)	89.62 % (7.91 ↓)
Caltech101	VP	**96.26 %**	80.70 % (15.56 ↓)	73.90 % (22.36 ↓)
	ILM-VP	**95.45 %**	80.53 % (14.92 ↓)	71.43 % (24.02 ↓)
OxfordPets	VP	**92.75 %**	60.86 % (31.89 ↓)	49.06 % (43.69 ↓)
	ILM-VP	**91.25 %**	58.79 % (32.46 ↓)	46.17 % (45.08 ↓)
SVHN	VP	**95.06 %**	89.30 % (5.76 ↓)	89.21 % (5.85 ↓)
	ILM-VP	**94.51 %**	89.18 % (5.33 ↓)	89.25 % (5.26 ↓)
GTSRB	VP	**91.43 %**	90.86 % (0.57 ↓)	90.73 % (0.70 ↓)
	ILM-VP	**91.06 %**	90.67 % (0.39 ↓)	88.79 % (2.27 ↓)
DTD	VP	**54.36 %**	36.38 % (17.98 ↓)	31.70 % (22.66 ↓)
	ILM-VP	**54.04 %**	40.48 % (13.56 ↓)	30.70 % (23.30 ↓)
EuroSAT	VP	**99.98 %**	98.03 % (1.95 ↓)	97.60 % (2.38 ↓)
	ILM-VP	**98.26 %**	97.78 % (0.98 ↓)	97.36 % (0.90 ↓)

Table 2. **Visual Prompting on Compressed CLIP**. Performance comparison of VP [4] and ILM-VP [7] on compressed CLIP ViT-L models across 7 datasets

where the transfer performance dips almost 45% in the 4x compressed model.

This outcome highlights how model compression negatively impacts not just the vision-only model transfer using visual prompting studied in the previous sections, but also the efficacy in downstream tasks in vision-language transformer-based models.

3.2.3 Quantized Models

We evaluate the performance of quantized Vision Transformer models, specifically comparing a full precision (32-bit) DeiT-T and Swin-T models with each of it's 4, 3, and 2-bit VVTQuantized DeiT-T and Swin-T model. [66].

In Figure 5, the overall trend indicates that the full precision DeiT-T model generally outperforms the quantized models across all datasets. Notably, at certain data budgets, the 4-bit DeiT-T model manages to match the performance of its full precision counterpart. The quantized models demonstrate clear performance differences as the quantization level increases. Similar trends are observed on additional datasets, as illustrated in the supplementary material.

In Figure 6, a similar trend is observed to that of the DeiT-T model, where the full precision model generally outperforms the quantized model. However, the gap between the 2-bit quantized model and others is smaller compared to what is seen in DeiT-T (more than 10% in some datasets like OxfordPets and DTD). This suggests nuanced dynamics in the efficacy of quantization methods across different model architectures.

In comparison to the ResNet-50 models pruned with AC/DC and RigL in the previous section, we observe that

the differences between a full precision and 4-bit quantized DeiT-T model are generally closer than those observed in sparsified ResNet-50 models. While sparsified models exhibit similar performance, in the quantization setting, as the level of quantization increases, the performance deviation among quantized models widens, especially for DeiT-T model. Quantized models and Specified AC/DC and RigL models demonstrate distinct behavioural deviations.

3.2.4 Cross Modal Reprogramming Results

In the Cross Modal Reprogramming setting, we extend our study to compare Dense Full Precision vision models against Sparse and Quantized vision models for NLP classification tasks.

The results in Table 3 reveal that Dense models generally outperform Sparse models, with the Splice DNA dataset being the only exception where the Sparse model outperforms the Dense one. Intriguingly, the Full-Precision model and its 3-bits Quantized counterpart show similar performance across most tasks. We provide 4-bit and 2-bit results in Supplementary material which further support these findings. In cross-model reprogramming to NLP tasks, we observe consistent good performance across different bit precision quantization.

Table 3. **Cross Modal Reprogramming Accuracy.** Comparison of Dense Full Precision, Sparse, and Quantized Vision Models for NLP Classification Tasks.

	Dense		*Sparse*		*Quantized*	
Task	Resnet-34	Resnet-50	Resnet-34	Resnet-50	Deit 32-bit	Deit 3-bit
Yelp	91.01 ± 1.21	90.93 ± 1.07	87.58 ± 1.63	86.03 ± 3.21	91.83 ± 1.41	92.08 ± 1.48
IMDB	79.90 ± 2.97	80.72 ± 3.75	68.46 ± 3.1	69.0 ± 0.52	77.86 ± 2.41	79.58 ± 2.64
AG	91.01 ± 0.62	90.6 ± 0.86	84.23 ± 0.99	85.87 ± 1.77	91.67 ± 0.88	92.48 ± 0.99
DBPedia	94.04 ± 1.50	94.12 ± 2.01	85.38 ± 0.86	78.19 ± 3.07	96.90 ± 1.35	96.65 ± 1.26
Splice	80.96 ± 0.57	74.84 ± 7.11	90.69 ± 2.97	81.37 ± 3.63	78.35 ± 8.70	71.57 ± 3.8

We provide an expanded table of Cross Modal Reprogramming results for quantized models, now including 4 bit and 2 bit in Table 4. The 4 bit and 3 bit quantized Deit models are generally able to keep similar performance to the full precision Deit model. The 2 bit model performs worse than the 32 bit model.

3.3. Calibration Analysis

Although there has been an increase in studies that explore the performance [17, 40, 48] and fairness [37, 38, 61] of sparse models, only recently have studies been aimed at studying the calibration of pruned models [2, 44]. Yet, how calibration varies under different levels and methods of compression, such as pruning and quantization, is largely unexplored. This is also the case, especially for transfer via VP. To this end, in this section, we analyze the calibration trends of pruned and quantized models at varying levels of

Table 4. **Cross Modal Reprogramming Accuracy.** Comparison of Quantized Vision Models for NLP Classification Tasks.

	Quantized			
Task	Deit 32-bit	Deit 4-bit	Deit 3-bit	Deit 2-bit
Yelp	91.83 ± 1.41	91.58 ± 1.21	92.08 ± 1.48	89.3 ± 2.6
IMDB	77.86 ± 2.41	78.68 ± 0.35	79.58 ± 2.64	79.0 ± 1.59
AG	91.67 ± 0.88	91.67 ± 1.47	92.48 ± 0.99	89.34 ± 1.21
DBPedia	96.90 ± 1.35	-	96.65 ± 1.26	95.18 ± 0.28
Splice	78.35 ± 8.70	76.96 ± 6.77	71.57 ± 3.8	61.76 ± 1.36

compression transferred by VP. Specifically, we inspect the expected calibration error (ECE) [29] as a measure of transfer reliability under VP and how it varies with increasing compression levels. To the best of our knowledge, this is the first work to present an extensive study on the calibration of compressed models transferred using VP.

Although there have been novel formulations for calibration analysis that are suited for niche settings such as deep ensembles [42] and dropout training [22], the ECE remains an important measure of reliability and is appropriate for our relative comparisons. Furthermore, we only analyze calibration of models transferred under a full-data setting, as compressed models tend to overfit or 'memorize' the training examples in a few-shot setting [3, 27], thus producing predictions that are already overconfident, and a measure of reliability under these low data-volume settings would not be a representative comparison.

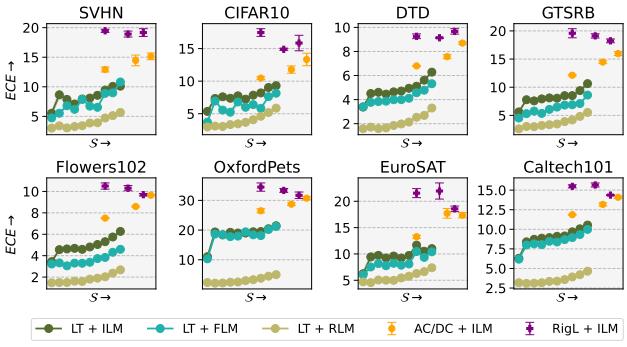

Figure 7. **Expected Calibration Error Analysis.** Comparison of ECE for LTH solutions of ResNet-50 models, and ResNet-50 models pruned by AC/DC and RigL transferred by ILM-VP across 8 datasets measured against increasing levels of model sparsity, starting from dense (left) to sparsest (right). Lower ECE is better.

The formulation [29] for ECE we use for our computations is given by,

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|M_b|}{N} \; |\mathrm{acc}(b) - \mathrm{conf}(b)|$$

where B is the total number of bins, $|M_b|$ is the number of predictions in bin b, N is the total number of samples, and $\text{acc}(b)$ and $\text{conf}(b)$ are the accuracy and the confidence of bin b respectively.

In Figure 7, we observe that the ECE of the dense transferred model (represented by the point on the left of each graph) is better than the sparse transferred models for all datasets, and this trend remains consistent regardless of the pruning technique used to compress the model.

Specifically for LTH solutions, the degradation in calibration performance on going from the dense model to the least-sparse transferred model is more pronounced for models that are transferred using ILM-VP and FLM-VP, while for RLM-VP we see a soft decline across most datasets. Furthermore, we see that with increasing levels of sparsity, the calibration of models continues to decline in all data-budget settings. The ECE for downstream performance on most datasets remains low around 10%, with OxfordPets being a notable exception for which even the least sparse model has a much worse ECE on transfer, surpassing 20%.

For models pruned by AC/DC and RigL and transferred by ILM-VP, we see that at the three levels of sparsity at which these models are evaluated, the ECE is significantly deteriorated compared to the dense counterpart. Note that the baseline for comparison for the AC/DC pruned models remains the dense ILM-VP calibration used for LTH since the underlying model remains ResNet-50 and these checkpoints are not progressively obtained. We also observe that the ECE of the AC/DC-pruned model is typically worse than that of the LTH solution when transferred using ILM-VP, and this effect is even more enhanced in the case of models pruned with RigL, which usually leads to models having relatively poorer calibration.

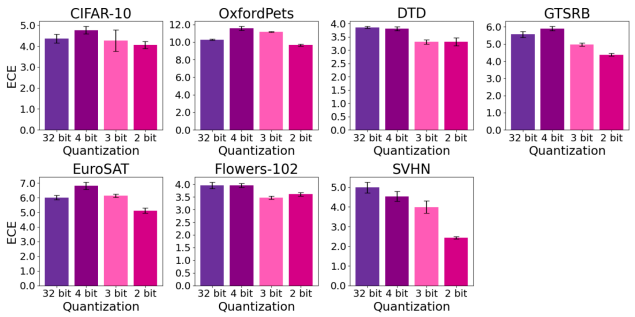

Figure 8. **DeiT ECE Analysis.** Comparison of ECE for DeiT quantized models across 7 datasets measured against varying precision of quantization, starting from full-precision (32-bit) to 2-bit. Lower ECE is better.

However, a distinct scenario unfolds when the compression method shifts to quantization. As depicted in Fig. 8 and Fig. 9, an increasing quantization rate, transitioning from full-precision (32-bit) to 2-bit, consistently reduces or maintains the same Expected Calibration Error (ECE). This discrepancy underscores the disparity in calibration

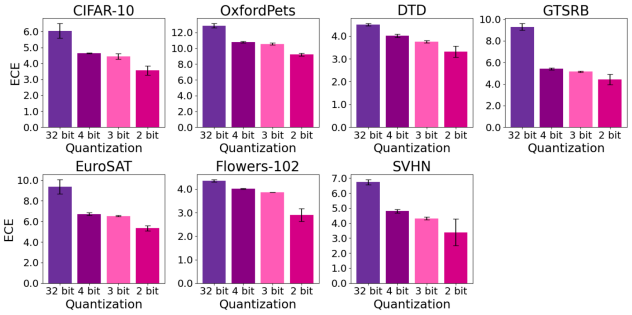

Figure 9. **Swin ECE Analysis.** Comparison of ECE for Swin quantized models across 7 datasets measured against varying precision of quantization, starting from full-precision (32-bit) to 2-bit. Lower ECE is better.

impact between models compressed via pruning and those compressed via quantization. To delve deeper into this phenomenon, we analyze the confidence values of both sparse and quantized models. Specifically, we examine two metrics: (a) Mean distance of the confidence distribution over classes for all incorrectly predicted samples compared to a uniform distribution, and (b) Mean confidence value for the correct class over all correctly predicted samples. We chose to measure the distance from the uniform distribution for incorrectly predicted samples because an ideal classifier should exhibit high uncertainty for incorrect predictions, indicating low confidence in assigning an incorrect label.

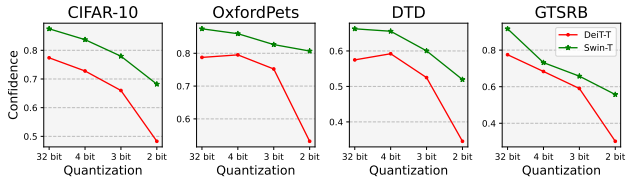

Figure 10. **Quantized Model Mean Confidence of Correct Prediction Analysis.** Observation of Mean confidence of the correct class for all accurate predictions for the DeiT-T and Swin-T models across full (32-bit) precision to 2-bit.

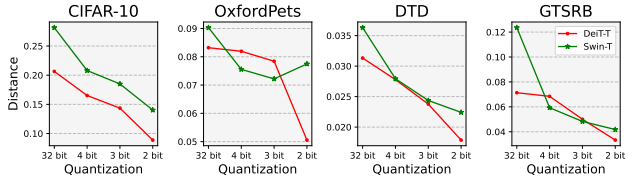

Figure 11. **Quantized Model Mean Distance to Uniform Distribution Analysis.** Observation of Mean of KL Divergence from the confidence distribution of incorrect model predictions to the uniform distribution for DeiT-T and Swin-T models across full (32-bit) precision to 2-bit.

In the case of quantized models, as depicted in Fig. 10 and Fig. 11, we observe an optimal pattern: as accuracy declines from the full-precision (32-bit) to the 2-bit variant, the mean confidence of correctly predicted class labels consistently decreases. Simultaneously, the mean distance of the confidence distribution from the uniform distribution for incorrectly classified samples also decreases monotonically, indicating a higher level of uncertainty when the overall model performance deteriorates.

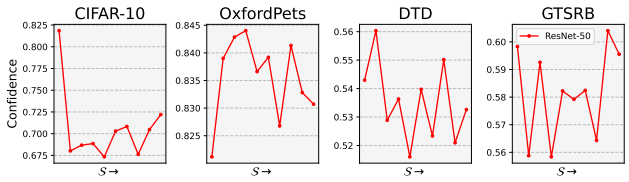

Figure 12. **Sparse Model Mean Confidence of Correct Prediction Analysis.** Mean Confidence of Correct Class for accurately predicted samples via ResNet-50 lottery ticket sparse models at different levels of sparsity for varying downstream target datasets. All results were analyzed via ILM-VP mode of transfer.

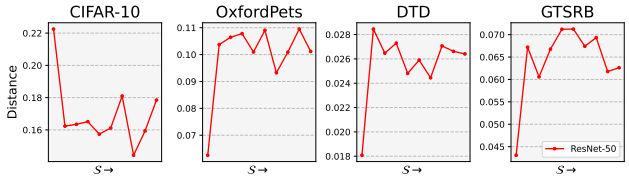

Figure 13. **Sparse Model Mean Distance to Uniform Distribution Analysis.** Mean of KL Divergence Distance between the confidence distribution of incorrectly classified samples and uniform distribution for sparse lottery ticket variants of ResNet-50 at varying levels of sparsity. All results were analyzed via ILM-VP mode of transfer.

However, when considering sparse models (LT), the optimal trend observed in quantized variants no longer persists. Specifically, as illustrated in Fig. 12, the confidence associated with the correct class label for accurately predicted samples fluctuates with an increasing rate of sparsity. More significantly, as sparsity intensifies, the distance of the confidence distribution for incorrectly classified samples from the uniform distribution either rises or fluctuates around the original dense model's values, indicating a higher level of overconfidence.

Given that Expected Calibration Error (ECE) is a metric representing the mean absolute difference between accuracy and corresponding confidence values, the observed trend in confidence values for sparse models elucidates why they encounter an increase in ECE. On the contrary, the more optimal trend observed in the case of quantization generally results in a lower or comparable ECE to that of the full dense, full-precision model.

4. Conclusion

In this work, we present an extensive study on the transfer of models using visual prompting methods on downstream classification tasks on the axis of model compression and low data volume. Our findings on a large number of datasets using pruned and quantized models from various vision-based architectures, as well as vision-language transformer-based architectures suggest the existence and universality of hidden cost in downstream performance drop when using visual prompting. We further show the detrimental impact of model sparsity on the calibration of the transferred model across models pruned by various techniques, which usually worsens with an increasing level of sparsity. Furthermore, on the contrary, we demonstrate that quantization as a method of compression does not exhibit the same negative impact on calibration as attributed in the case of sparse (pruned) models. With the rapid advances in vision(-language) foundation models, visual prompting can be a crucial technique for downstream task adaptation, similar to the indispensable role of text prompting for large-language models, and our results provide important insights and motivations into the future design of prompting-friendly model compression methods. Following our analysis into the hidden costs associated with model compression, our aim is to extend our work by using influence estimation [18] to characterize the positive or negative contribution of certain data sub-populations to transfer performance under the visual prompting regime, and also on a wider range of compression techniques.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 1

[2] Viplove Arora, Daniele Irto, Sebastian Goldt, and Guido Sanguinetti. Quantifying lottery tickets under label noise: accuracy, calibration, and complexity. *arXiv preprint arXiv:2306.12190*, 2023. 2, 3, 6

[3] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 6

[4] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 1, 2, 4, 5

[5] Rémi Bernhard, Pierre-Alain Moellic, and Jean-Max Dutertre. Impact of low-bitwidth quantization on the adversarial robustness for embedded neural networks. In *2019 International Conference on Cyberworlds (CW)*, pages 308–315. IEEE, 2019. 2

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[7] Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, and Sijia Liu. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19133–19143, 2023. 2, 3, 4, 5

[8] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. *arXiv preprint arXiv:2202.10629*, 2022. 1, 2

[9] Tianlong Chen, Zhenyu Zhang, Pengjun Wang, Santosh Balachandra, Haoyu Ma, Zehao Wang, and Zhangyang Wang. Sparsity winning twice: Better robust generalization from more efficient training. *arXiv preprint arXiv:2202.09844*, 2022. 2

[10] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[12] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale, 2022. 2, 3

[13] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[15] Gamaleldin F Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. *arXiv preprint arXiv:1806.11146*, 2018. 1, 2

[16] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning*, pages 2943–2952. PMLR, 2020. 3, 4

[17] William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*, 2022. 6

[18] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020. 8

[19] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018. 2, 3

[20] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023. 2, 3

[21] Yonggan Fu, Ye Yuan, Shang Wu, Jiayi Yuan, and Yingyan Lin. Robust tickets can transfer better: Drawing more transferable subnetworks in transfer learning. *arXiv preprint arXiv:2304.11834*, 2023. 3

[22] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 6

[23] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574, 2019. 3

[24] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *ArXiv*, abs/2103.13630, 2021. 3

[25] James Gilles. *The lottery ticket hypothesis in an adversarial setting*. PhD thesis, Massachusetts Institute of Technology, 2020. 2

[26] Brunno F Goldstein, Sudarshan Srinivasan, Dipankar Das, Kunal Banerjee, Leandro Santiago, Victor C Ferreira, Alexandre S Nery, Sandip Kundu, and Felipe MG França. Reliability evaluation of compressed deep learning models. In *2020 IEEE 11th Latin American Symposium on Circuits & Systems (LASCAS)*, pages 1–5. IEEE, 2020. 2

[27] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 6

[28] Micah Gorsline, James Smith, and Cory Merkel. On the adversarial robustness of quantized neural networks. In *Proceedings of the 2021 on Great Lakes Symposium on VLSI*, pages 189–194, 2021. 3

[29] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 3, 6

[30] Masafumi Hagiwara. A simple and effective method for removal of hidden units and weights. *Neurocomputing*, 6(2): 207–218, 1994. 3

[31] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 2, 3

[32] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks, 2015. 2, 3

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[34] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE international geoscience and remote sensing symposium*, pages 204–207. IEEE, 2018. 3

[35] Dan Hendrycks and Mantas Mazeika. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*, 2022. 3

[36] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021. 3

[37] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019. 2, 6

[38] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020. 2, 6

[39] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, number 1288, 2013. 3

[40] Eugenia Iofinova, Alexandra Peste, Mark Kurtz, and Dan Alistarh. How well do sparse imagenet models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12266–12276, 2022. 3, 6

[41] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[42] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 6

[43] Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605, 1990. 2, 3

[44] Bowen Lei, Ruqi Zhang, Dongkuan Xu, and Bani Mallick. Calibrating the rigged lottery: Making all tickets reliable. *arXiv preprint arXiv:2302.09369*, 2023. 2, 3, 6

[45] Chenhao Li, Qiang Qiu, Zhibin Zhang, Jiafeng Guo, and Xueqi Cheng. Learning adversarially robust sparse networks via weight reparameterization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8527–8535, 2023. 2

[46] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022. 3

[47] Ruofeng Li. *Calibration Analysis of Structured Pruning Methods and Cascade Classifier Design Based on Pruned Networks*. McGill University (Canada), 2022. 2

[48] Jianyi Lin. Sparse models for machine learning. In *Engineering Mathematics and Artificial Intelligence*, pages 107–146. CRC Press. 6

[49] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Zahra Atashgahi, Lu Yin, Huanyu Kou, Li Shen, Mykola Pechenizkiy, Zhangyang Wang, and Decebal Constantin Mocanu. Sparse training via boosting pruning plasticity with neuroregeneration. *Advances in Neural Information Processing Systems*, 34:9908–9922, 2021. 3

[50] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*, pages 2736–2744, 2017. 2

[51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[52] Paarth Neekhara, Shehzeen Hussain, Jinglong Du, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley. Cross-modal adversarial reprogramming. In *WACV*, 2022. 4

[53] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 3

[54] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 3

[55] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2, 3

[56] Alexandra Peste, Eugenia Iofinova, Adrian Vladu, and Dan Alistarh. Ac/dc: Alternating compressed/decompressed training of deep neural networks. *Advances in neural information processing systems*, 34:8557–8570, 2021. 3, 4

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5

[58] Hadi Salman, Andrew Ilyas, Logan Engstrom, Sai Vemprala, Aleksander Madry, and Ashish Kapoor. Unadversarial examples: Designing objects for robust vision. *Advances in Neural Information Processing Systems*, 34:15270–15284, 2021. 1

[59] Dachuan Shi, Chaofan Tao, Ying Jin, Zhendong Yang, Chun Yuan, and Jiaqi Wang. Upop: Unified and progressive pruning for compressing vision-language transformers, 2023. 3, 5

[60] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. arxiv 2020. *arXiv preprint arXiv:2012.12877*, 2020. 2, 3

[61] Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 35:17652–17664, 2022. 2, 6

[62] Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *International Conference on Machine Learning*, pages 9614–9624. PMLR, 2020. 1, 2, 3, 4

[63] Hsi-Ai Tsao, Lei Hsiung, Pin-Yu Chen, Sijia Liu, and Tsung-Yi Ho. Autovp: An automated visual prompting framework and benchmark. *arXiv preprint arXiv:2310.08381*, 2023. 1

[64] Bindya Venkatesh, Jayaraman J Thiagarajan, Kowshik Thopalli, and Prasanna Sattigeri. Calibrate and prune: Improving reliability of lottery tickets through prediction calibration. *arXiv preprint arXiv:2002.03875*, 2020. 2, 3

[65] Ria Vinod, Pin-Yu Chen, and Payel Das. Reprogramming pretrained language models for protein sequence representation learning. *arXiv preprint arXiv:2301.02120*, 2023. 1, 2

[66] Zhiqiang Shen Xijie Huang and Kwang-Ting Cheng. Variation-aware vision transformer quantization. *arXiv preprint arXiv:2307.00331*, 2023. 2, 3, 5

[67] Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. Voice2series: Reprogramming acoustic models for time series classification. In *International conference on machine learning*, pages 11808–11819. PMLR, 2021. 1, 2

[68] 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, 2022. 2, 3

[69] Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. Fairness reprogramming. *Advances in Neural Information Processing Systems*, 35:34347–34362, 2022. 1, 2

[70] Hao Zhou, José Manuel Álvarez, and Fatih Murat Porikli. Less is more: Towards compact cnns. In *European Conference on Computer Vision*, 2016. 2

[71] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[72] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017. 4