Exploring Intrinsic Dimension for Vision-Language Model Pruning

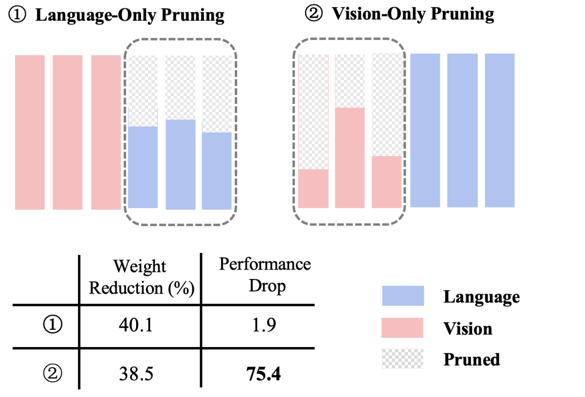
Hanzhang Wang [1] * Jiawen Zhang [1] Qingyuan Ma [1]

Abstract

The intrinsic dimension (ID) represents the minimum dimension needed to describe data on a lower-dimensional manifold within high-dimensional spaces. Network pruning aims to reduce the complexity of high-dimensional networks while minimizing performance trade-offs. This symmetry motivates the exploration of ID as a metric for effective pruning. For vision-language models, we investigate whether different modalities exist on separate manifolds, indicating varying complexity and prunability. We empirically study ID variations in large-scale vision-language pre-trained models and examine the contributions of different modalities to model prunability. We propose a layer importance metric based on ID, which can conveniently integrate with current metrics and enhance performance in vision-language model pruning. The experimental results show a high correlation between ID and modality prunability. Visual representations are more sensitive and crucial to model performance, while language representations are more robust and offer greater prunability. Our findings suggest an asymmetric pruning strategy for vision and language modalities, guided by the ID metric. The code is available at https://github.com/Nofear18/ID_VL_Pruning

1. Introduction

In the pursuit of advancing large-scale vision-language models, a question arises: How does the representation formed by billions of parameters relate to the underlying complexity of the data they represent? While the current large-scale models are expansive, encompassing several billion parameters, their sheer size does not inherently translate to superior data abstraction or effective generalization. Networks that are over-parameterized often fall prey to overfitting, capturing incidental training noise instead of the essential data distribution (Srivastava et al., 2014; Liebenwein et al., 2021).



	Weight Reduction (%)	Performance Drop
①	40.1	1.9
②	38.5	**75.4**

Figure 1. Comparison of model weight reduction and performance for language-only and visual-only pruning strategies. At approximately 40% pruning ratio, the language-only and vision-only strategies result in a 1.9 and 75.4 drop in the CIDEr metric, respectively.

This study focuses on the interplay between intrinsic and extrinsic dimensionality, a key aspect in understanding the efficiency and effectiveness of large-scale pre-trained models. The extrinsic dimension, typically the network's architecture size (e.g., 512, 1024), often stands in stark contrast to its intrinsic dimension, which varies based on different inputs and tasks. The intrinsic dimension provides a more essential description of data, being significantly smaller than its extrinsic counterpart. Studies (Li et al., 2018; Brown et al., 2023) have demonstrated the variability of the intrinsic dimension (ID) across several datasets for image classification, indicating a disparity between a network's theoretical architectural capacity and the complexity of the data it represents. In small-scale neural networks, it has been demonstrated that less than 1% of the subspace dimensions can retain 90% of the original model's performance (Li et al., 2018). These findings suggest a potential for significant pruning of parameters to achieve more effective representations.

In assessing weight importance for pruning, leading strate-

[1] School of Computer Engineering and Science, Shanghai University. Correspondence to: Hanzhang Wang <hanzhang.mon.wang@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

gies primarily consider two elements: the distribution (Han et al., 2015; Zhu & Gupta, 2017; Alford et al., 2019) of the weights and their gradients (Molchanov et al., 2019; Xiao et al., 2019; Sanh et al., 2020), which compare all weights across the entire network. However, these methods often fail to consider the hierarchical structure of the network and its role in data abstraction and generalization. After pruning, the hierarchical representations of the network are disrupted, where previously each layer independently served as an abstraction of the data. Maintaining hierarchical integrity is necessary to ensure that the pruned network retains the capability to reconstruct the original data, essential for effective representation and generalization.

Visual and language modalities differ fundamentally in their structural forms and the intricacy of the information they embody. Visual content is often immediate and information-rich, capturing complex scenes and emotions in a single frame. In contrast, textual content is abstract, requiring the decoding of language to extract meaning.

In the architecture of multi-modal models such as CLIP (Radford et al., 2021), the vision and language components may have similar extrinsic dimensions—768×3072 for vision layers and 1024×4096 for language layers. Yet, their impact on model performance is markedly different, reflecting the distinct semantic complexities they encapsulate. As Figure 1 demonstrates, pruning 40% of weights from the vision or language parts alone results in a significant discrepancy in performance impact, with the CIDEr metric dropping by 75.4 for vision and only 1.9 for language. Upop (Shi et al., 2023) also observe that the language modality is more easily prioritized for pruning through an adaptive search-based pruning algorithm.

Inspired by the notion of mapping data from higher, extrinsic dimensions to more fundamental, intrinsic ones. Our study extends it to the pruning of vision-language models. We focus on the intrinsic dimensions of visual and language representations and their interplay within low-dimensional manifolds, employing IDs as a metric for evaluating the importance of weights in pruning vision-language pre-training models. The main contributions of this work are as follows.

- We empirically investigate the intrinsic dimensions in vision-language pre-train and their pruning models, providing a detailed analysis of the correlation between ID geometric properties and modal prunability. IDs of visual modality are varied with a hunchback shape, ranging from 20 to 450. In contrast, IDs of language modality are uniform with a lower range from 5 to 30. Notably, language representations are more robust with higher prunability, while vision representations are more sensitive and have a greater impact on overall performance.
- We propose to utilize IDs to measure the layer-wise importance of pruning, which can be easily cooperated with other metrics. The experiment results demonstrate that utilizing the ID as an indicator for weight pruning yields superior performance across multiple tasks. Compared to the full model, the 40% pruned model has only a 1.9 drop on the CIDEr metric.

2. Related Work

2.1. Weight Importance Metrics

The importance score can be broadly classified into two categories: gradient-based importance scores and sensitivity-based importance scores. Let $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ represent all parameters in a network, and $\boldsymbol{\theta}_{i^*} = [0, \ldots, 0, \theta_i, 0, \ldots, 0] \in \mathbb{R}^d$ the pruned parameters. The importance score of $\boldsymbol{\theta}$ is denoted as $S(\theta)$.

Magnitude-based method quantifies the importance of a weight by its absolute value, denoted as $S(\theta_i) = |\theta_i|$. Although magnitude-based scoring facilitates efficient pruning, this simple metric may not comprehensively reflect a parameter's contribution to the model's output. There are instances where weights with relatively small magnitudes can still exert a substantial influence on the overall model performance (Yang & Liu, 2022; Molchanov et al., 2019).

Han et al. (2015) implement magnitude-based pruning in deep networks by removing weights below a certain threshold. Building on this, Zhu & Gupta (2017) and Alford et al. (2019) show that pruned, large models can outperform smaller, dense models with similar memory usage. Renda et al. (2020) further improves re-training by adjusting the learning rate alone. Additionally, Li et al. (2016) uses this approach to select filters for pruning.

Sensitivity-based method quantifies the significance of a parameter by evaluating the loss incurred in the model output when the parameter is zeroed. The score $S(\theta_i)$ be defined as $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}\left(\boldsymbol{\theta} - \boldsymbol{\theta}_{i^*}\right)$. If removing a parameter leads to a substantial increase in model loss, it indicates higher sensitivity, and consequently higher importance.

Liang et al. (2021) uses this criterion and verifies that a subnetwork within the model does most of the heavy lifting, and the remaining weights can all be pruned away. Molchanov et al. (2019) describes two variations of this method using the first and second-order Taylor expansions to approximate a filter's contribution. Similarly, Sanh et al. (2020) proposes movement pruning which uses a deterministic first-order weight pruning method that is adaptive to a pre-trained model with fine-tuning. Recently, Zhang et al. (2022) points out that the sensitivity is not reliable during pruning. They propose to resolve this issue by sensitivity smoothing and uncertainty quantification.

2.2. The Intrinsic Dimensionality of Vision and Language

Vision. Current work on intrinsic dimension mainly focuses on the unimodal, particularly the visual modality. Ansuini et al. (2019) investigates the ID profile of three common CNN-based pre-trained representations and finds the hunchback shape of the ID variation across the layers. Muratore et al. (2022) observes similar first-expansion-then-reduction of object representations along the rat homolog of the ventral stream. Pope et al. (2021) estimates the ID of several popular datasets and finds that common natural image datasets have very low intrinsic dimensions relative to the high number of pixels in the images. ID is used to study the semantic complexity of synthetic images by GAN (Pope et al., 2021; Horvat & Pfister, 2022; Barannikov et al., 2021), which allows actively manipulating the ID by controlling the image generation process. Brown et al. (2023) empirically verify the hypothesis of the union of manifolds in common image datasets and find that the data lies on a disconnected set with varying IDs. Amsaleg et al. (2017) and Ma et al. (2018) use the local ID to characterize the adversarial robustness of attacked visual regions and find that the LID increases along with the increasing noise in adversarial perturbations.

Language. Compared with the large number of ID studies on visual information, including datasets and representations, there are fewer studies on the ID characteristics of language modality. Fine-tuning of the large language model, BERT (Kenton & Toutanova, 2019) and RoBERTa (Liu et al., 2019) are analyzed from the ID perspective in Aghajanyan et al. (2020). Both theoretical and empirical explanations have been provided, pointing to a low-dimensional reparameterization that is as effective in fine-tuning as the full parameter space. Kvinge et al. (2023) focuses on the prompts for text-to-image generation. It demonstrates that prompt variations affect the ID of model layers in distinct ways. Bottleneck layers, instead of latent layers, correlate with prompt perplexity and intrinsic dimension. Tulchinskii et al. (2023) finds that the average intrinsic dimensionality of fluent texts in natural language hovers around the value of 7 to 9 for human-generated texts, while the average ID of AI-generated texts for each language is around 1.5 or even lower. The clear statistical separation enables a simple classifier to distinguish human-generated and AI-generated texts.

Aghajanyan et al. (2020) presents a methodology for reducing parameters in the fine-tuning of large language models through the utilization of ID. In the realm of Neural Architecture Search (NAS), He et al. (2023) employs ID to gauge similarities among architectures, revealing that ID-based characterizations offer enhanced space separability and superior performance ranking scores compared to gradient-based methods. Moreover, Ankner et al. (2022) posits that as the ID of a neural network increases, its prunability correspondingly decreases. Based on these findings, we propose a hypothesis that incorporating ID into the process of model pruning could effectively indicate the significance of parameters to a certain extent.

3. Method

3.1. The TwoNN Algorithm

The TwoNN algorithm (Facco et al., 2017), is utilized to estimate the intrinsic dimension (ID) of the representations produced by each layer of a pre-trained model. This estimation is achieved by analyzing the distances between each point and its nearest and second nearest neighbors, denoted as r_1 and r_2, respectively. The ratio of these distances, μ, inherently less than 1, increases in value as the ID escalates. It is important to note that μ adheres to a Pareto distribution, $Pa(d+1)$, where d represents the intrinsic dimension. The likelihood of a sample set of this distribution, $\mu = (\mu_1, \mu_2, \ldots, \mu_N)$, is given by the formula:

$$P(\mu|d) = d^N \prod_{i=1}^{N} \mu_i^{-d-1}. \tag{1}$$

This formula underpins the linear regression approach to solving for ID, sidestepping global distribution assumptions by focusing on constant density around each point.

The TwoNN algorithm is recognized for its effectiveness, grounded in its use of distance ratios to the nearest neighbors for estimating the intrinsic dimension (ID) of data. Unlike methods that depend on assumptions about data density or the smoothness of the data manifold, TwoNN's approach is more direct and less prone to inaccuracies when applied to high-dimensional datasets. By relying solely on local information from nearest neighbors, it circumvents the challenges associated with data scaling and rotation. This attribute positions the TwoNN algorithm as a more straightforward alternative to Maximum Likelihood Estimation (MLE) (Levina & Bickel, 2004) and other methods that require detailed assumptions about the data's geometric properties or density.

3.2. Estimating the IDs of Vision and Language Representations

We follow the implements of Ansuini et al. (2019) to sequentially estimate the ID of each layer's representations, conducting a statistical analysis of ID across different layers. Each layer's ID is estimated separately, utilizing the MSCOCO dataset (Lin et al., 2014) to ensure consistency in the evaluation framework.

Given the computational complexity of the TwoNN algorithm, which operates with a time and space complexity

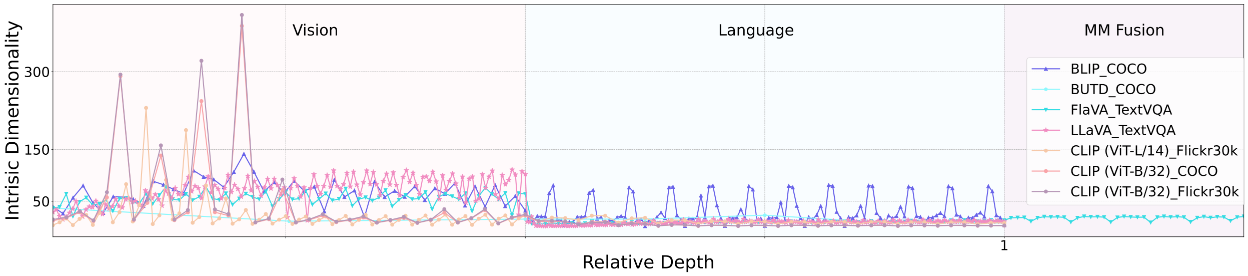


Figure 2. ID variations are analyzed across various datasets for the BUTD (Anderson et al., 2018), BLIP (Li et al., 2022), CLIP (Radford et al., 2021), FlaVA (Singh et al., 2022), and LLaVA (Liu et al., 2023) pre-training models. FlaVA incorporates a separate multi-modal fusion model, while other models typically feature independent vision and language models. ID estimation is implemented using TwoNN with 2,000 samples.

of $O(n_D^2)$, the choice of dataset size, n_D, is pivotal for the effectiveness of estimation. Facco et al. (2017) empirically recommends selecting a sample size approximately ten times the anticipated intrinsic dimension. Aligning with this guideline, our implementation employs 2,000 samples, balancing between computational feasibility and the need for robust statistical representation.

We employ the BLIP model (Li et al., 2022), CLIP (Radford et al., 2021), FlaVA (Singh et al., 2022), LLaVA (Liu et al., 2023), and BUTD (Anderson et al., 2018) as surrogates to explore the characteristics of multi-modal representations. We estimate the fully connected layer representations of these models on various datasets including MSCOCO (Lin et al., 2014), TextVQA (Singh et al., 2019), and Flickr30k (Plummer et al., 2017). These pretrain models vary in architectures, modal fusion manner, and modal scale.

The BUTD model (Anderson et al., 2018) features a single vision Transformers model coupled with a language model using LSTMs, establishing a basic framework for image-text interactions. The CLIP model (Radford et al., 2021) adopts a dual-encoder architecture, employing either a Vision Transformer or a ResNet for image and a Transformer for text. It utilizes contrastive learning trained on a large dataset of image-text pairs to improve the alignment between visual and textual modalities. BLIP (Li et al., 2022) extends CLIP by integrating a multimodal Transformer that merges visual and textual representations. It includes a text-generating decoder and uses both contrastive and captioning losses, aiming to enhance tasks such as image captioning and visual question answering. The FlaVA model (Singh et al., 2022) introduces an additional multimodal fusion module to further fuse vision and language representations. LlaVA (Liu et al., 2023) combines the vision model of CLIP ViT-L/14 with the large language model of Vicuna (Chiang et al., 2023), integrating advanced vision processing with superior language understanding capabilities.

Figure 2 illustrates the ID variations in layer-wise representations, revealing distinct distribution patterns across vision and language modalities. Specifically, the visual modality demonstrates a diverse distribution, typically resembling a 'hunchback' shape with a wide range of values from 20 to 450. This observation aligns with findings of CNN-based visual representations from Ansuini et al. (2019); Muratore et al. (2022). In contrast, the language modality displays a more uniform and cyclical distribution, consistent across various sizes and complexities of language models, with values generally ranging from 5 to 30. Notably, peaks within the BLIP language model correspond to the key and value layers of cross-modality attention, rather than in pure language representations. The IDs of multimodal layers (typically the BLIP and FlaVA models) show a periodic pattern similar to that of language, yet their values fall between those of the visual and language modalities.

3.3. Iterative Pruning with Intrinsic Dimension

Algorithm 1 Iterative Pruning with Intrinsic Dimension

```
Input: Neural Network Model $M$, Total Training Epochs $K$, Final Pruning Ratio $P_{final}$
Estimate ID for each layer in Model $M$
for $epoch = 1$ to $K$ do
  if $epoch$ is between 2 and $K - 1$ then
    Update $P_{current}$ towards $P_{final}$
    Compute and prune weights based on $S \times ID$
  end if
  Train Model $M$ for the current epoch
end for
```

Iterative pruning is conducted across K epochs, following the process described in Algorithm 1, with K set to 5 in our main experiments. Within this framework, the model is initially trained without pruning to establish a baseline. Starting from the second epoch, the pruning ratio $P_{current}$ is progressively adjusted towards the final pruning target P_{final}, employing a cubic schedule (Sanh et al., 2020) to modulate the pruning intensity at each step.

Table 1. Comparison with predominant weight importance metrics: Magnitude (Zhu & Gupta, 2017), Gradient (Molchanov et al., 2019), and PLATON(Zhang et al., 2022). Pruning results on MSCOCO datasets image caption task with pruning ratios at 40%, 80%, 90%, and 95%.

	CIDEr				BLEU@4				SPICE			
Full Model	**133.3**				**39.7**				**23.8**			
Pruning Ratio	40%	80%	90%	95%	40%	80%	90%	95%	40%	80%	90%	95%
Magnitude	**131.5**	84.9	35.3	17.6	39.0	27.7	15.1	10.3	**23.7**	16.2	8.1	4.8
Magnitude*ID	131.0	102.7	56.5	26.5	38.9	33.2	21.3	12.9	23.6	19.1	12.3	6.9
Gradient	108.0	47.2	29.0	20.8	33.9	18.6	13.9	11.2	20.1	10.3	6.7	5.2
Gradient*ID	110.4	69.7	40.4	24.9	34.0	23.1	16.9	12.2	20.6	14.0	9.5	6.0
PLATON	130.7	124.1	93.1	**39.5**	39.0	37.6	30.3	**16.7**	23.5	22.6	17.9	**9.1**
PLATON*ID	131.4	**129.2**	**106.4**	39.0	**39.2**	**39.1**	**33.9**	16.4	**23.7**	**23.3**	**19.8**	8.9

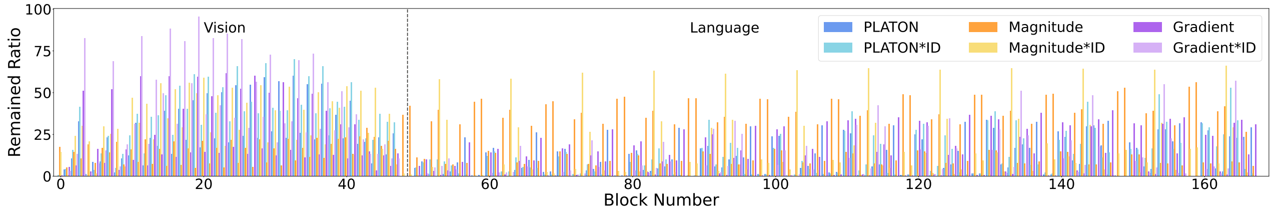


Figure 3. Layer-wise pruning ratio comparison for Magnitude, Gradient, and PLATON metrics, and their corresponding scores when multiplied by ID. The entire model pruning ratio is 80%.

Importance scores for pruning are dynamically updated every 20 steps within each epoch, incorporating the intrinsic dimension (ID) of each layer. By multiplying the original importance score S by the ID, the pruning process prioritizes that pruning decisions are informed by both the weight importance and the layer representation complexity.

4. Experimental Results

Setup. All of our experiments are conducted on 4 NVIDIA GeForce GTX 3090 GPUs using PyTorch. To evaluate our metric, we compare it with several leading pruning metrics. All metrics are tested under the same conditions, details please see Appendix. The metrics for comparison are:

Magnitude (Zhu & Gupta, 2017) targets weights with the smallest absolute value for pruning, based on the premise that smaller weights contribute less to the model's output. **Gradient** (Molchanov et al., 2019) prunes weights that result in the smallest change in loss, suggesting that such weights have a minimal impact on the model's learning. **PLATON** (Zhang et al., 2022) considers the magnitude, gradient, and uncertainty of the gradient in its importance evaluation. It offers a smooth pruning criterion across mini-batches by accounting for the variability of the gradient.

4.1. Image Captioning

Datasets. We evaluate image captioning performance using the MSCOCO dataset (Lin et al., 2014), encompassing 80 object and 91 stuff categories with a standard split of 118K training images and 5K images each for validation and testing, each image accompanied by 5 human-annotated captions. Evaluation metrics include CIDEr (Vedantam et al., 2015), BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE (Lin, 2004), and SPICE (Anderson et al., 2016).

Results. Table 1 shows the pruning results across various sparsity levels. The PLATON*ID method outperforms other pruning metrics in all evaluated metrics, especially at higher pruning ratios (80%, 90%, and 95%). Comparing Magnitude and Gradient methods with their ID-enhanced metrics, it is evident that incorporating ID generally boosts pruning performance. This integration not only preserves but can enhance model performance, even with significant reductions in network size. However, the performance gains from ID diminish as the pruned model's performance nears that of the unpruned model.

Figure 3 illustrates how incorporating ID with pruning metrics such as Magnitude, Gradient, and PLATON affects layer-wise pruning ratios. Notably, when ID is considered, visual layers are pruned less extensively, highlighting their

Table 2. NLVR task: Accuracy of pruned models with various metrics on NLVR2 dataset.

Full Model	**83.4**	
Pruning Ratio	40%	90%
Magnitude	52.9	51.1
Magnitude*ID	79.8	51.1
Gradient	51.1	51.1
Gradient*ID	75.0	51.1
PLATON	81.4	72.7
PLATON*ID	**81.5**	**73.1**

Table 3. Image classification task: Pruned model performance using different metrics, compared with ID integration on CIFAR-100 dataset.

Full Model	**87.6**	
Pruning Ratio	40%	95%
Magnitude	83.6	44.5
Magnitude*ID	**84.3**	45.1
Gradient	66.0	25.7
Gradient*ID	64.7	26.0
PLATON	83.0	63.6
PLATON*ID	82.8	**63.7**

Table 4. Image classification task: Pruned model performance using different metrics, compared with ID integration on ImageNet dataset.

Full Model	**79.3**	
Pruning Ratio	40%	80%
Magnitude	77.8	65.6
Magnitude*ID	78.4	66.5
Gradient	40.7	29.2
Gradient*ID	41.5	28.8
PLATON	78.4	70.5
PLATON*ID	**78.5**	**71.4**

critical role in the overall vision-language representation. These results support our finding that layers with higher intrinsic dimensions require more conservative pruning.

Conversely, integrating ID enables more aggressive pruning in language layers, indicating a lower ID and a less critical role in network performance. Specifically, the PLATON*ID metric facilitates increased pruning in language layers, reflecting their diminished importance in overall performance.

It is noteworthy that the comparison results from Figure 3 indicate that ID, or the model shape inferred from layer-wise IDs, is not the only factor influencing model performance. PLATON exhibits the best pruning performance, followed by Magnitude. Despite this, the shapes of their layer-wise pruning rate curves, whether ID is incorporated or not, show substantial differences. Conversely, while the curves of PLATON and Gradient are more similar, the performance of the models they represent varies significantly.

4.2. Natural Language for Visual Reasoning

Dataset. We evaluate the visual reasoning task using the NLVR2 dataset (Suhr et al., 2018), which addresses a binary classification problem: determining whether a textual description accurately corresponds to a pair of images. This dataset includes over 107,000 instances, challenging models to effectively align and reason across both visual and textual modalities. Performance is measured by accuracy, reflecting the proportion of instances correctly predicted.

Results. Table 2 demonstrates the sensitivity of the NLVR2 task to reductions in model size. At a 40% pruning ratio, models pruned without Intrinsic Dimensionality (ID) approximate random guessing, with accuracy slightly above 50%. However, integrating ID with the Magnitude and Gradient metrics significantly boosts performance, yielding accuracies of 79.8% and 75.0%, respectively.

In extreme pruning scenarios (90% pruning ratio), while the Magnitude and Gradient metrics integrated with ID do not show performance gains, remaining at an accuracy of 51.1%, the PLATON metric's effectiveness is marginally improved by ID, increasing from 72.7% to 73.1% accuracy.

4.3. Image Retrieval

Dataset. The Flickr30k dataset comprises over 30,000 images collected from the online photo-sharing website Flickr, each annotated with five different textual descriptions by humans. The sentences, averaging between 10 to 20 words, describe the scenes with varying levels of detail and complexity.

Results. Table 5 shows that the PLATON*ID model significantly improves performance across all recall metrics at a 40% pruning ratio on the Flickr30k dataset. Both Text-to-Image and Image-to-Text retrieval tasks see enhanced results, with notable increases in Recall@1, Recall@5, and Recall@10, demonstrating the effectiveness of integrating ID with the PLATON metric.

Table 5. Performance of the CLIP model on the Flickr30k dataset at 40% pruning ratio.

Model	T2I			I2T		
	R@1	R@5	R@10	R@1	R@5	R@10
PLATON	66.2	87.2	92.5	63.0	87.0	92.6
PLATON*ID	**71.7**	**91.5**	**95.6**	**64.6**	**88.2**	**93.2**

4.4. Image Classification

To rigorously evaluate the effectiveness of ID in pruning, we conduct experiments on visual models, where the disparity in ID across layers is more pronounced compared to language models, thereby validating its effectiveness more distinctly. Our baseline visual model is the ViT-B/16 model (Dosovitskiy et al., 2022), which is tested on the image classification task.

Dataset. We utilize the CIFAR-100 (Krizhevsky, 2009) and ImageNet-1k (Russakovsky, 2015) datasets to benchmark image classification performance. CIFAR-100 contains 60,000 images across 100 classes, and ImageNet-1k offers a larger scale with over 592,000 images in 1,000 categories, providing a comprehensive challenge for recognition models.

Results. Tables 3 and Table 4 demonstrate that integrating Intrinsic Dimensionality (ID) with pruning metrics typically improves performance, especially when used with the PLATON metric. On CIFAR-100, incorporating ID slightly enhances the accuracy of the Magnitude metric at a 40% pruning ratio and maintains it at 95%. For the Gradient metric, the effect of ID is mixed; it slightly reduces accuracy on CIFAR-100 at a 40% pruning ratio but increases it at both pruning ratios on ImageNet. When combined with ID, the PLATON metric either maintains or modestly improves accuracy across both datasets and at all pruning ratios. Notably, at a 40% pruning ratio on ImageNet, PLATON*ID marginally surpasses the accuracy of the full model, and at 80%, it exceeds the performance of PLATON alone, highlighting the effectiveness of ID in enhancing the pruning process for high-performance models.

A comprehensive comparison of the results from Table 2, Table 3, and Table 4 reveals that ID is ineffective for models with very poor performance, and offers limited improvement for models that already perform well. We infer that for ID to substantially contribute, the model should possess basic representational capabilities and a substantial pruning ratio to provide enough retraining opportunity for ID to significantly boost its effectiveness.

5. Discussion

5.1. Prunability of Vision and Language

To understand the different behaviors of vision and language modalities in response to pruning, we prune each modality separately and together in the full model using the PLATON metric across various ratios (from 20% to 95%). We evaluate performance using the CIDEr metric.

Figure 4 demonstrates that models pruned solely in the visual modality experience a greater reduction in performance compared to models pruned only in the language modality, which maintain relatively stable CIDEr scores even at 95% pruning. This observation suggests that language models may have a higher prunability threshold, possibly due to their stable and lower ID distribution. The most substantial performance reductions are observed in visual-only pruned models, consistent with the trends in the full models.

Table 6 assesses the impact of pruning on different metrics. The language model, pruned at an 80% ratio with the PLATON*ID metric, shows minimal performance loss, indicating that language representations are generally more tolerant to pruning. This observation suggests that the integration of ID effectively identifies non-essential weights. In contrast, vision-only models pruned with PLATON/ID experience significant drops in performance, highlighting their dependency on maintaining certain critical weights. Language models maintain steady performance under both PLATON and PLATON*ID metrics, which implies a substantial presence of expendable weights in pre-trained language models, enabling them to recover quickly after fine-tuning, even following extensive pruning. This disparity emphasizes the evaluation of weight importance, especially for visual models. Moreover, these findings suggest that the ID metric, particularly when integrated as a layer importance multiplier, provides a better gauge for maintaining essential representations.

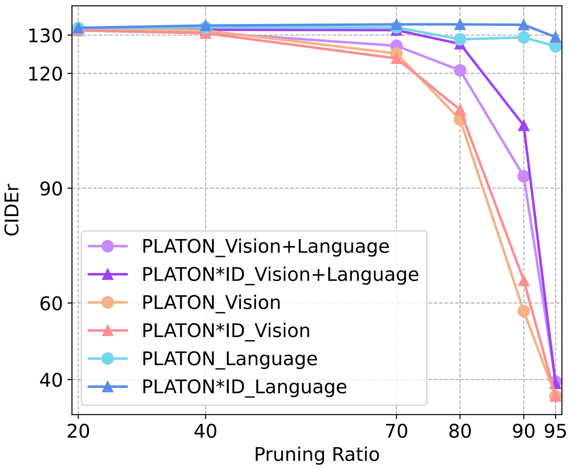


Figure 4. CIDEr performance at pruning ratios from 20% to 95% for vision, language, and vision+language using PLATON and PLATON*ID metrics.

Table 6. When pruning the vision model (V), language model (L), and the entire model (V+L) at an 80% pruning ratio, we compare the performance of pruned models across multiple metrics, demonstrating the impact of ID on weight importance evaluation.

Model	Prune	C	B@4	M	R	S
Full Model	-	**133.3**	**39.7**	**30.9**	**60.0**	**23.8**
PLATON	V	107.9	33.4	27.5	55.3	20.4
	L	128.9	39.0	30.0	59.3	23.0
	V+L	124.1	37.6	29.6	58.5	22.6
PLATON*ID	V	110.5	33.8	27.9	55.8	20.7
	L	**132.8**	**39.7**	**30.6**	**59.9**	**23.8**
	V+L	129.2	39.1	30.2	59.4	23.3
PLATON/ID	V	101.1	31.5	26.0	54.0	19.4
	L	128.7	38.7	28.9	59.3	22.9
	V+L	75.0	25.2	22.3	48.9	15.1

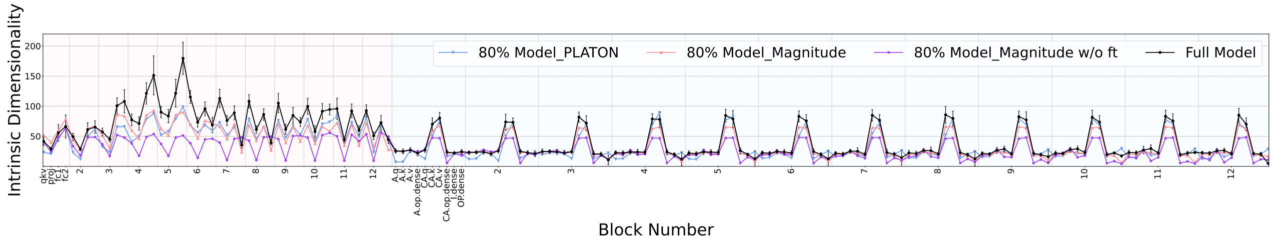


Figure 5. ID changes induced by various pruning methods and training strategies at an 80% pruning ratio. Pruning generally lowers ID across all layers, while fine-tuning tends to increase ID as performance improves. Models with the lowest performance show the smallest IDs and flatten variation in ID within the vision modality.

5.2. How Pruning Changes the IDs of Vision and Language?

While it is intuitive that pruning distorts the original representations, there is debate about how this distortion influences intrinsic dimensionality (ID). Muratore et al. (2022) argue that pruning luminosity and contrast information in visual representations increases the ID value. In contrast, Ankner et al. (2022) contend that the prunability of a neural network decreases as the ID increases. We empirically explore this issue by pruning models using different metrics, including Magnitude and PLATON, both with and without fine-tuning.

Magnitude. A baseline method proposed by Zhu & Gupta (2017) that prunes parameters with small magnitudes over five iterative epochs.

Magnitude w/o FT. Similar to the Magnitude method but without any fine-tuning, set up to examine the ID changes solely due to pruning.

PLATON. An approach proposed by Zhang et al. (2022) that incorporates both gradient and uncertainty in its pruning metric, aimed at exploring ID changes with a superior performance.

Prune both modalities. Table 7 and Figure 5 present the performance and Intrinsic Dimensionality (ID) of models pruned at an 80% ratio across various metrics. We observe that magnitude pruning without fine-tuning results in the lowest ID values, which correlate with a significant drop in performance. In contrast, the PLATON method achieves significantly better performance compared to magnitude pruning, even though their ID values are similar in most layers.

Interestingly, pruning generally reduces ID values across most layers. However, an exception is noted in the PLATON method for certain layers of the language model, where post-pruning IDs are higher than those of the unpruned model. On the other hand, the magnitude without fine-tuning model consistently exhibits the lowest ID values across layers, which correlates with its poorer performance relative to the full model.

Table 7. Model performance using different weight importance metrics when pruning ratio is 80%. C, B, M, R, and S denote CIDEr, BLEU, METEOR, ROUGE_L, and SPICE.

Model	C	B@4	M	R	S
Full Model	133.3	39.7	30.9	60.0	23.8
Mag w/o FT	0.4	9^{-10}	1.7	9.2	0.0
Mag	77.6	25.8	22.7	49.3	15.5
PLATON w/o FT	0.2	1^{-8}	2.6	18.2	0.0
PLATON	124.1	37.6	29.6	58.5	22.6

A detailed layer-wise comparison reveals variations in ID values between the magnitude and PLATON models. In the vision layers, the magnitude model shows higher ID values in the initial blocks, while PLATON surpasses it after the initial layers. In the language layers, both models have similar ID values, though PLATON reaches a higher peak ID. Overall, we derive the following observations:

1) Pruning generally reduces IDs across all layers, and there is a positive correlation between ID values and model performance.

2) In vision-language pre-training models, the vision component is more affected by pruning than the language component, as indicated by a stronger correlation between ID values and performance, regardless of the pruning strategy.

3) Although ID values positively correlate with model performance, this relationship is not strictly linear. It appears that the highest ID within the network may be a more robust predictor of performance than previously suggested. Specifically, while Ansuini et al. (2019) propose that the ID of the final layer predominantly dictates performance in unpruned models, our findings suggest that this might not fully apply to pruned models.

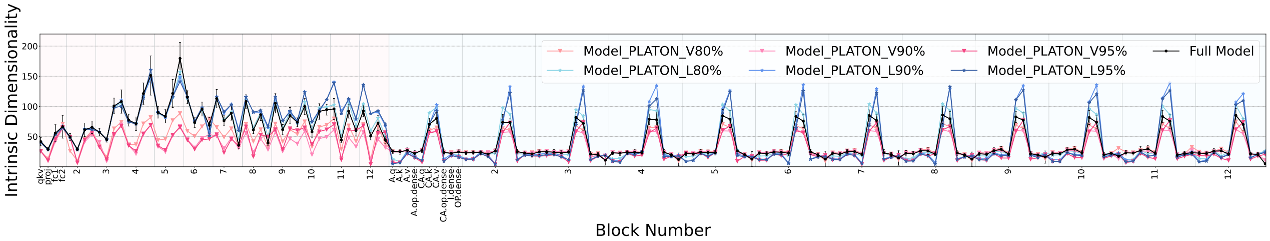


Figure 6. When one modality is pruned (red for vision, blue for language) at 80%, 90%, and 95% ratios, its effect on the ID changes of both modalities.

Prune single modality. Further investigations are directed towards the impact of modality-specific pruning on the Intrinsic Dimensionalities (IDs). Figure 6 illustrates the ID variations when pruning only the vision (red lines) or language models (blue lines) at 80%, 90%, and 95% pruning ratios. The IDs of the unpruned full model are shown with a black line. Based on these comparisons, we make the following observations:

1) Vision-only pruning generally results in a decrease in ID across both the visual and language layers.

2) Language-only pruning leads to an increase in the ID of vision layers, including the key and value layers of cross-attention, while causing a decrease in the ID of pure language layers.

Assuming that the magnitude of ID correlates with the generalization capability of representations, we can infer several insights from these observations. Pruning visual layers not only distorts visual representations but also affects language representations, especially at higher pruning ratios.

In the case of language-only pruning, we observe a significant increase in ID in the later visual layers and key-value layers in cross-attention. This suggests that pruning the language model refines the network's ability to process visual information, potentially enhancing the efficiency of cross-modal interactions. Although pruning degrades pure language representation, it may improve its alignment with visual representation.

Moreover, the changes in ID due to language pruning may have reverse effects on the visual layers, as indicated by the upward trends in the later vision layers (blue lines). When language representation is refined, it may enhance the capacity of visual representation, potentially benefiting it through back-propagation.

6. Limitations

Intrinsic Dimensionality (ID) offers a potential method for determining the optimal number or inter-layer ratio of minimal dimensions necessary for neural network layers. However, achieving optimal performance with the dimensions specified by ID requires further study. In this work, we integrate Intrinsic Dimensionality (ID) with model pruning based on layer importance, but we do not explore training or fine-tuning strategies. Maintaining the original model performance with a very small number of parameters requires additional investigation, such as developing effective training strategies for parameter-efficient networks. Notably, Li et al. (2018) discusses this issue in small-scale fully connected (FC) and convolutional neural networks. Their findings suggest that it is possible to achieve a highly compressed model that retains over 90% of its original performance through random subspace training. Consequently, exploring effective training and initialization strategies that leverage ID remains a crucial area for future research, especially for large-scale pruned models.

7. Conclusion

This study investigates the intrinsic dimensionality (ID) of a pre-trained multi-modal model to inform weight pruning strategies. We find that visual modalities exhibit a wide range of ID values, suggesting varied layer importance, while language modalities show more consistent variations. The experimental results demonstrate that ID effectively assesses layer significance, thereby improving pruning. The language modality is robust to pruning despite possessing many redundant weights, whereas the vision modality, though sensitive, is crucial for model performance. These observations offer a manifold geometry perspective to interpret vision and language representations, potentially providing insights for optimizing architecture and training strategies for vision-language models.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62206167), Natural Science Foundation of Shanghai (23ZR1423500), Shanghai University Youth Talent Sailing Plan, and Shanghai Technical Service and Computing Center of Science and Engineering at Shanghai University.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Alford, S., Robinett, R., Milechin, L., and Kepner, J. Training behavior of sparse neural network topologies. In *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–6. IEEE, 2019.

Amsaleg, L., Bailey, J., Erfani, S., Furon, T., Houle, M. E., Radovanovic, M., and Vinh, N. X. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2017.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. Spice: Semantic propositional image caption evaluation. In *ECCV*, pp. 382–398, 2016.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

Ankner, Z., Renda, A., Dziugaite, G. K., Frankle, J., and Jin, T. The effect of data dimensionality on neural network prunability. *arXiv preprint arXiv:2212.00291*, 2022.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems 32*, pp. 6391–6402, 2019.

Banerjee, S. and Lavie, A. Meteor: an automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Barannikov, S., Trofimov, I., Sotnikov, G., Trimbach, E., Korotin, A., Filippov, A., and Burnaev, E. Manifold topology divergence: a framework for comparing data manifolds. *Advances in Neural Information Processing Systems*, 34:7294–7305, 2021.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. Verifying the union of manifolds hypothesis for image data. In *ICLR*, 2023.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2022.

Facco, E., d'Errico, M., Rodriguez, A., and Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

He, X., Yao, J., Wang, Y., Tang, Z., Cheung, K. C., See, S., Han, B., and Chu, X. Nas-lid: efficient neural architecture search with local intrinsic dimension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7839–7847, 2023.

Horvat, C. and Pfister, J.-P. Intrinsic dimensionality estimation using normalizing flows. In *NeurIPS 2022*, 2022.

Kenton, J. D. M.-W. C. and Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2, 2019.

Krizhevsky, A. e. a. Learning multiple layers of features from tiny images. 2009.

Kvinge, H., Brown, D., and Godfrey, C. Exploring the representation manifolds of stable diffusion through the lens of intrinsic dimension. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Levina, E. and Bickel, P. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

Li, C., Farkhoor, H., Liu, R., and Yosinski, J. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

Li, H., Kadav, A., Durdanovic, I., Samet, H., and Graf, H. P. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Liang, C., Zuo, S., Chen, M., Jiang, H., Liu, X., He, P., Zhao, T., and Chen, W. Super tickets in pre-trained language models: From model compression to improving generalization. *arXiv preprint arXiv:2105.12002*, 2021.

Liebenwein, L., Baykal, C., Carter, B., Gifford, D., and Rus, D. Lost in pruning: The effects of pruning neural networks beyond test accuracy. *Proceedings of Machine Learning and Systems*, 3:93–138, 2021.

Lin, C.-Y. Rouge: a package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll'ar, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

Molchanov, P., Mallya, A., Tyree, S., Frosio, I., and Kautz, J. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11264–11272, 2019.

Muratore, P., Tafazoli, S., Piasini, E., Laio, A., and Zoccolan, D. Prune and distill: similar reformatting of image information along rat visual cortex and deep neural networks. In *36th Conference on Neural Information Processing Systems*, 2022.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Plummer, A. B., Wang, L., Cervantes, M. C., Caicedo, C. J., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, pp. 74–93, 2017.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, pp. 6391–6402, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Renda, A., Frankle, J., and Carbin, M. Comparing rewinding and fine-tuning in neural network pruning. *arXiv preprint arXiv:2003.02389*, 2020.

Russakovsky, O. e. a. Imagenet large scale visual recognition challenge. In *International journal of computer vision*, volume 115, pp. 211–252. Springer, 2015.

Sanh, V., Wolf, T., and Rush, A. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems*, 33:20378–20389, 2020.

Shi, D., Tao, C., Jin, Y., Yang, Z., Yuan, C., and Wang, J. Upop: Unified and progressive pruning for compressing vision-language transformers. *arXiv preprint arXiv:2301.13741*, 2023.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Piontkovskaya, I., Nikolenko, S., and Burnaev, E. Intrinsic dimension estimation for robust detection of ai-generated texts. *arXiv preprint arXiv:2306.04723*, 2023.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Xiao, X., Wang, Z., and Rajasekaran, S. Autoprune: Automatic network pruning by regularizing auxiliary parameters. *Advances in neural information processing systems*, 32, 2019.

Yang, C. and Liu, H. Channel pruning based on convolutional neural network sensitivity. *Neurocomputing*, 507: 97–106, 2022.

Zhang, Q., Zuo, S., Liang, C., Bukharin, A., He, P., Chen, W., and Zhao, T. Platon: Pruning large transformer models with upper confidence bound of weight importance. In *International Conference on Machine Learning*, pp. 26809–26823. PMLR, 2022.

Zhu, M. and Gupta, S. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.

A. Implementation Details

Hyper-parameters. We use a cubic pruning schedule similar to Sanh et al. (2020); Zhang et al. (2022) for the experiments in rows 1-4 of Table 8. This schedule includes initial warm-ups, t_i, and final warm-ups, t_f, defined as:

$$r^{(t)} = \begin{cases} r^{(0)} & \text{if } 0 \leq t < t_i \\ r^{(T)} + \left(r^{(0)} - r^{(T)}\right)\left(1 - \frac{t - t_i - t_f}{T - t_i - t_f}\right)^3 & \text{if } t_i \leq t < T - t_f \\ r^{(T)} & \text{otherwise} \end{cases} \tag{2}$$

where $t_i = i \times l$, $t_f = f \times l$, and l is the length of the training dataloader.

For the experiments in rows 5 and 6 of Table 8, we perform 50 steps of iterative pruning, setting a target pruning rate at each step and evaluating the model performance. All experiments use the AdamW optimizer (Loshchilov & Hutter, 2018), with additional hyperparameters detailed in Table 8.

Table 8. Hyper-parameter of our all experiments. Rows 1-4 correspond to progressive pruning, and rows 5-6 correspond to one-shot pruning.

Task	Model	Dataset	Batch Size	Epochs	i	f	Initial LR	Weight Decay
Image Captioning	BLIP	COCO	20	5	1	1	$1e-5$	0.00
Visual Reasoning	BLIP	NLVR2	10	10	0	5	$3e-5$	0.05
Image Classification	ViT-B/16	CIFAR-100	128	50	0	20	$5e-4$	0.05
Image Classification	ViT-B/16	ImageNet	128	15	0	10	$5e-4$	0.05
Image Captioning	BLIP	COCO	20	0	-	-	$1e-5$	0.00
Image Retrieval	CLIP	Flickr30k	2	0	-	-	$1e-5$	0.05