1. Introduction and motivation

In the main paper, we analyze the potential problems of traditional model compression and knowledge distillation methods. Inspired by the principle of minimally invasive surgery, we propose a brand-new model compression method named Minimally Invasive Surgery. MIS learns the principal features from a pair of dense and compressed models in a contrastive manner. We prove that MIS changes the learning effectiveness ratio and the probability distribution between easy and hard learning objects from information entropy and Bayes perspectives. With the comparison and ablation experiments, we show the success of MIS relies on learning the inherent discrepancy between the representation capacities of the dense and the compressed model, and the discrepancy introduced by hardware acceleration restrictions between two compressed models. With MIS, we can compress the models for various tasks into efficient forms and can get considerable acceleration in general-purpose GPUs.

The motivation of MIS method combines two main points in the contribution list.

- MIS is designed to have better performance than traditional knowledge distillation and the other network compression methods.
- MIS needs to provide the end-to-end compression for neural networks to meet the specific hardware acceleration requirements.

When making the investigation of the network compression methods aiming at model sparsity, we can divide them into two categories, i.e., coarse-grained sparsity and fine-grained sparsity. For the coarse-grained sparsity like filter-sparsity and channel-sparsity, the regular sparse pattern is easy to achieve acceleration on general-purpose processors because it is equivalent to a smaller dense model. For the fine-grained sparsity, the acceleration on general-purpose hardware is very limited due to the irregular sparse pattern caused by the compression methods. Because there are many papers focus on the coarse-grained sparsity (e.g., filter pruning researches), so the main focus of the proposed MIS method is on the fine-grained sparse model.

A100 GPU [19] has the new feature to support the fine-grained structured sparsity by enforcing through a 2:4 sparse matrix definition that allows two non-zero values in every four-entry vector, as shown in Figure 1. Due to the well-defined structure of the matrix, it can be compressed efficiently and reduce memory storage and bandwidth by almost 2X. The sparse Tensor Cores can exploit 2:4 structured sparsity to double the compute throughput of standard Tensor Core operations for neural networks. So if we can make tiny changes on the irregular fine-grained sparse pattern like minimally invasive surgery, and match the 2:4 fine-grained sparse requirements. Then we can fully utilize this feature to provide extra acceleration to the fine-grained sparse model.

In this Appendix, we will provide some supplementary materials and more experimental results for the proposed MIS algorithm.

2. Experimental results

In the experiments section of the main paper, we have investigated these issues:

- MIS effectiveness and performance across most of the comment networks and applications.
• Quantitative comparison with state-of-the-art methods.
• Ablation experiments.
• Acceleration on general-purpose hardware.

In the Appendix, we will provide more results with different parameters settings. For the experiments in this section, we choose PyTorch [21] to implement all algorithms. Most of the training and fine-tuning experimental results are obtained with V100 GPU clusters [18]. The acceleration performance results are obtained with A100 GPU clusters [19] to fully utilize its Tensor Core [20] support for fine-grained structured sparsity and irregularly-compressed models. Because V100 and A100 GPUs could provide much larger math throughput than FP16 than FP32 data type, we also combine MIS with the mixed-precision training [17] provided by APEX\(^5\) to compress the models into a more hardware-efficient format. So all the accuracy results reported by MIS are using FP16 as the default data type. All the reference algorithms use the default data type provided in public repositories. (Almost all use FP32 except where noted.)

2.1. Effectiveness experiments for classification task

To evaluate the effectiveness of the MIS method on the image classification task, we take the ResNet-50 [8], ResNeXt-101 [28] and MobileNet-V2 [23] from TorchVision\(^6\) as the experiment target models.

In the main paper, the loss adjustment parameters among the surgical prediction loss (\(\alpha\)), the healthy-surgical distillation loss (\(\beta\)) and the recovered-surgical distillation loss (\(\gamma\)) apply 1, 10, 50, respectively. More results with different adjustment parameters can refer to Table 1. (The variance is within ±0.17 for Top-1 accuracy, and ±0.15 for Top-5 accuracy with different random seeds.)

The original sparse models serve as \(M_\text{OR}\) trained with the public Distiller library\(^7\) [33]. \(\ast\)-PRE represents the pre-trained model. \(\ast\)-FINE represents the fine-grained sparse model obtained by adopting a gradual pruning technique (AGP) to sparsify the model during the training process\(^4\) [31]. \(\ast\)-BLK represents the block-grained [32] sparse model, \(\ast\)-SUR represents the fine-grained [6] sparse model by applying pruning and splicing in a dynamical manner, \(\ast\)-SNIP represents the single-shot pruned [13] model by analyzing the connection sensitivity. In this experiment, MIS does not use the ground truth label provided by ImageNet [4] dataset. It takes the predicted label from \(M_\text{OR}\) to calculate the surgical prediction loss.

\(^1\)https://github.com/NVIDIA/apex.
\(^2\)https://github.com/pytorch/vision.
\(^3\)https://github.com/NervanaSystems/distiller.
\(^4\)Notice some of the sparse ResNet-50 models and all of the sparse ResNeXt-101 models have higher accuracy than the pre-trained dense models provided by TorchVision.

<table>
<thead>
<tr>
<th>Models</th>
<th>Original Accuracy</th>
<th>Recovered Model Accuracy</th>
<th>Parameter Sparsity</th>
<th>Shared Memory</th>
<th>Random Seed</th>
<th>Recovered vs. Healthy</th>
<th>Recovered vs. Fake Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50-FINE 90%</td>
<td>78.088</td>
<td>59.868</td>
<td>90%</td>
<td>80%</td>
<td>10</td>
<td>90%</td>
<td>80%</td>
</tr>
<tr>
<td>ResNet50-FINE 95%</td>
<td>79.289</td>
<td>60.089</td>
<td>95%</td>
<td>80%</td>
<td>10</td>
<td>95%</td>
<td>80%</td>
</tr>
<tr>
<td>MobileNet-V2-FINE 75%</td>
<td>68.752</td>
<td>50.672</td>
<td>75%</td>
<td>80%</td>
<td>10</td>
<td>75%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Table 1. MIS effectiveness on image classification task.

2.2. Effectiveness experiments for detection task

To evaluate the effectiveness of the MIS on the detection task, we take the Faster R-CNN [22], RetinaNet [14], Mask R-CNN [7] from Detectron\(^5\), and SSD [16] from NVIDIA repository\(^8\) as the experiment target models. In the main paper, the loss adjustment parameters among the surgical prediction loss (\(\alpha\)), the healthy-surgical distillation loss (\(\beta\)) and the recovered-surgical distillation loss (\(\gamma\)) apply 1, 10, 15, respectively. More results with different adjustment parameters can refer to Table 2. (The variance is within ±0.20 for average precision, and ±0.24 for average recall with different random seeds.)

The original sparse models serve as \(M_\text{OR}\) are compressed with AGP method and trained with the Distiller library\(^3\). R50, R101 and X101 in the brackets represent the ResNet-50, ResNet-101 and ResNeXt-101 models served as the backbone of the detection networks. \(\text{AR}\) and \(\text{AP}\) represent the average precision and average recall metrics. In this experiment, MIS uses the ground truth info provided by COCO [15] dataset.

\(^5\)https://github.com/facebookresearch/detectron2.
2.3. Effectiveness experiments for translation task

To evaluate the effectiveness of the MIS on the translation task, we take the GNMT [27] from NVIDIA repository and Transformer [25] from Fairseq as the experiment target models. In the main paper, the loss adjustment parameters among the surgical prediction loss ($\alpha$), the healthy-surgical distillation loss ($\beta$) and the recovered-surgical distillation loss ($\gamma$) apply 1, 1.5, 3, respectively. More results with different adjustment parameters can refer to Table 3. (The variance is within ±0.13 for **PSNR**, and ±0.045 for **SSIM** with different random seeds.) The representative super-resolution outputs are shown in Figure 2.

The original sparse models serve as $M_h$ are compressed with the pruning method [3]. WMT14 En-Ge and WMT16 En-Ge in the brackets represent the WMT14 and WMT16 English-German dataset, respectively. In this experiment, MIS uses the ground truth info provided by WMT datasets.

### 2.4. Effectiveness experiments for super resolution

To evaluate the effectiveness of the MIS on the super resolution task, we take the SRResNet [12] as the experiment target model. In the main paper, the loss adjustment parameters among the surgical prediction loss ($\alpha$), the healthy-surgical distillation loss ($\beta$) and the recovered-surgical distillation loss ($\gamma$) apply 1, 1.5, 3, respectively. More results with different adjustment parameters can refer to Table 4. (The variance is within ±0.13 for **PSNR**, and ±0.045 for **SSIM** with different random seeds.) The representative super-resolution outputs are shown in Figure 2.

The original sparse models serve as $M_h$ are compressed with the pruning method [9]. SRResNet is trained on the English-German dataset, respectively. 

Figure 2. Representative super resolution results with enlargements of boxed areas (The Recovered Model and Surgical Model are compressed to 50% sparse level).
DIV2K dataset [1]. The DIV2K validation images, as well as Set5 [2] and Set14 [29] datasets are used to report deployment quality. In the super resolution task, image quality is often evaluated by two metrics: Peak Signal-to-Noise Ratio (PSNR) [10] and Structural Similarity (SSIM) [26].

2.5. Ablation experiments and insights

2.5.1 More accurate healthy model

We change the healthy model with a more accurate one to verify whether it can further improve the effect of MIS. We use the pre-trained ResNeXt-101 from TorchVision as the healthy model. The results are shown in Table 5.

<table>
<thead>
<tr>
<th>Model</th>
<th>Algorithm</th>
<th>Sparsity Ratio</th>
<th>Top-1 (%)</th>
<th>Top-5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>Baseline</td>
<td>0%</td>
<td>76.130</td>
<td>92.862</td>
</tr>
<tr>
<td></td>
<td>BLK</td>
<td>70%</td>
<td>76.452</td>
<td>92.990</td>
</tr>
<tr>
<td></td>
<td>AGP</td>
<td>7%</td>
<td>76.496</td>
<td>93.080</td>
</tr>
<tr>
<td></td>
<td>KD</td>
<td>7%</td>
<td>75.950</td>
<td>92.710</td>
</tr>
<tr>
<td></td>
<td>RKD</td>
<td>7%</td>
<td>75.474</td>
<td>93.124</td>
</tr>
<tr>
<td></td>
<td>CRD</td>
<td>7%</td>
<td>76.432</td>
<td>93.190</td>
</tr>
<tr>
<td></td>
<td>MIS(unsup)</td>
<td>7%</td>
<td>75.910</td>
<td>92.650</td>
</tr>
<tr>
<td></td>
<td>MIS(sup)</td>
<td>7%</td>
<td>76.558</td>
<td>93.188</td>
</tr>
</tbody>
</table>

Table 5. MIS with more accurate healthy model.

From the results, we can conclude a more accurate healthy model can bring extra benefit in accuracy. It also proves that MIS can be used when dense and compressed models have different structures. This is not realizable for the model compression methods which rely on distillation from pure feature maps, like LIT [11].

2.5.2 Contribution of each component

In this experiment, we want to check the contribution of each component in MIS to the final model compression effect. Then we can have a deep insight into why MIS can outperform state-of-the-art methods. Apart from AGP and KD methods we have discussed, we also involve the Residual Knowledge Distillation [5] (RKD) and Contrastive Representation Distillation [24] (CRD) methods in the comparison. The results with different sparsity ratio are shown in Table 6. Unsupervised and Supervised in the brackets represent MIS does not use & use the ground truth info provided by ImageNet, respectively.

MIS introduces two distillation loss items to learn the inherent discrepancy between the representation capacities of the dense and the compressed model, and the discrepancy introduced by hardware acceleration restrictions between two compressed models. From the results, we can see these key differences from KD, RKD, and CRD contribute to the good effect of MIS.

We can also find even without the ground truth info from the training set, MIS can still achieve satisfactory accuracy.
2.5.3 Visualization

We apply the Class Activation Mapping (CAM) tool [30] to the healthy model $M_H$, the recovered model $M_R$ and the surgical model $M_S$ for ResNet-50. CAM can highlight the importance of the image region to the final prediction. The visualization results are shown in Figure 3.

![Visualizations of healthy, recovered, and surgical models](image)

Figure 3. Class activation mapping visualization. (The Recovered Model and Surgical Model are compressed to 80% sparse level).

For CAM, the red color highlights the “attention” area of each model. Though the surgical model is restricted by the hardware acceleration requirements, the CAMs of $M_H$, $M_R$ and $M_S$ all focus on the inherent features of the Malinois, red fox and face powder in the ground truth images, which leading to the right classification.

3. Conclusion and ethics statement

For the open-source community, our experimental observations and the proposed compression technique could be inspiring to the model compression field. Our study also provides good guidance for people who want to try the latest features for the newly announced A100 GPU.

Mobile applications performing object detection or super-resolution on the client to save bandwidth can benefit from simpler models. Using efficient models in the data centers can leave more resources available to train much more complex networks.

From the societal impact aspect, the neural models are widely used to daily tasks like autonomous driving, medical imaging, etc. Our proposed compression technique can bring beneficial impacts on various applications. So compressed models with higher deployment efficiency will help in pedestrian detection, emergency protection, medical analysis, and diagnosis. And eventually protecting people’s safety and saving more lives.
References

[18] NVIDIA. NVIDIA Tesla V100 GPU Architecture, 2017. 2
[19] NVIDIA. NVIDIA A100 Tensor Core GPU Architecture, 2020. 1, 2

