# Effect of CNN Compression on Gradient-based Explainability

**Anonymous Authors**[1]

## Abstract

This project investigates the intricate relationship between CNN compression, employing knowledge distillation, and its impact on gradient-based explainability. Focusing on the teacher model WRN-28-10 and student model Resnet-18, trained with similar accuracies, we employ objective metrics - average drop, coherency, and complexity, to assess changes in explainability. Subtle shifts in activation maps reveal nuanced differences in feature focus between the models. Notably, the student model exhibits a slightly lower average confidence drop and lower complexity, suggesting improved confidence in predictions, and more coherent feature retention. Our findings suggest that model compression, particularly with deep student models like ResNet-18, can enhance explainability metrics, prompting considerations for future investigations into shallower models and the evolving landscape of interpretability in compressed CNNs.

## 1. Introduction

### 1.1. Background

Convolutional Neural Networks (CNNs) are like superstars in computer vision—they excel at recognizing images, spotting objects, and breaking down scenes. Their knack for learning complex features automatically has made them a big deal in artificial intelligence. But, here's the catch: these high-flying CNNs are pretty heavy on the brainpower and memory, making them a tough fit for gadgets with limited resources or places with slow internet.

As the integration of AI into devices with restricted resources gains momentum in real-world applications, the imperative for efficient model deployment has intensified. Model compression techniques step into this arena by mitigating the size and computational demands of CNNs, a crucial endeavor for facilitating their deployment in edge devices, mobile applications, and contexts where resource efficiency is paramount.

Yet, amid the advantages of model compression, a growing concern has arisen regarding the potential compromise of interpretability and explainability in compressed models.

### 1.2. Motivation

The motivation behind investigating the impact of CNN compression on gradient-based explainability lies in the increasing adoption of model compression techniques for resource-efficient deployment. As models undergo compression through methods like knowledge distillation or pruning, comprehending their internal mechanisms becomes pivotal for ensuring trust, transparency, and accountability. While efficiency gains are paramount, preserving interpretability ensures that even compressed models remain comprehensible and trustworthy in their decision-making processes.

### 1.3. Objectives

This research addresses the critical intersection of CNN compression and gradient-based explainability, seeking to unravel how compression methods, particularly knowledge distillation, impact the interpretability of these neural networks in real-world applications. We try to quantitatively measure interpretability metrics for explainability methods and translate the findings into practical insights for deploying compressed CNNs, emphasizing the balance between compression and interpretability in real-world applications.

## 2. Related Work and Concepts

### 2.1. Understanding Explainability in ML

In the field of machine learning, there is no clear agreement on what "explainability" means, and this has resulted in different and sometimes unclear explanations. Achieving explainability in machine learning involves two main aspects: interpretability, which means making the model understandable to humans, and fidelity, which means accurately describing how the model works. Interpretability is further divided into clarity and simplicity, focusing on

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

providing clear and straightforward explanations for users. Fidelity includes ensuring the model is both accurate and covers the entire task it was designed for.

Explainability in machine learning not only promotes social acceptance, trust, and interaction with these systems but also contributes to safety and knowledge acquisition. Enhancing the explainability of systems increases safety by aiding in identifying faulty behavior and easing testing, auditing, and debugging. Additionally, it facilitates successful human interaction with machine learning systems, aligning their operation with intended goals.

### 2.2. Novel metrics for explainability

Regarding Convolutional Neural Networks (CNNs), which are central to many deep learning tasks, achieving explainability poses challenges due to their complex architecture and millions of parameters. Visualizing saliency maps and adopting methods like Class Activation Mapping (CAM) and Grad-CAM are instrumental in providing insights into CNN predictions. CAM, introduced by Zhou et al. (2016), enables object localization without requiring bounding box annotations and highlights crucial image regions influencing CNN decisions. Similarly, Grad-CAM, proposed by Selvaraju et al. (2020), aims to capture high-level semantics and detailed spatial information from deep convolutional layers. By computing gradients of scores with respect to activation maps and subsequent global average pooling, Grad-CAM generates class-discriminative visualizations, offering insights into significant image parts for classification decisions.

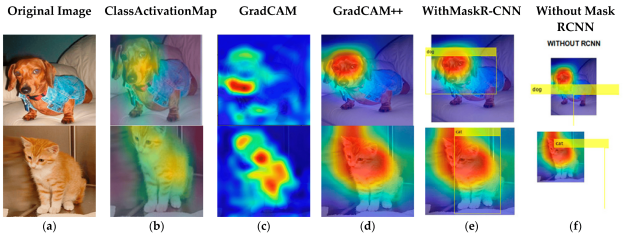


Figure 1: Visualisation of different explainability methods

Class Activation Mapping (CAM) methods have emerged as effective visualization tools, employing weighted averages of activation maps to provide insightful visual representations. However, evaluating and comparing CAM-based approaches necessitate comprehensive metrics to gauge the efficacy of explanation maps systematically.

$$CAM_c(x) = \text{ReLU}\left(\sum_{k=1}^{N_l} \alpha_k A_k\right)$$

A novel set of metrics specifically designed to quantify explanation maps using the activation mapping proposed in the paper (Poppi et al., 2021) is listed below. We would be using these metrics in the project for comparing two CAMs:

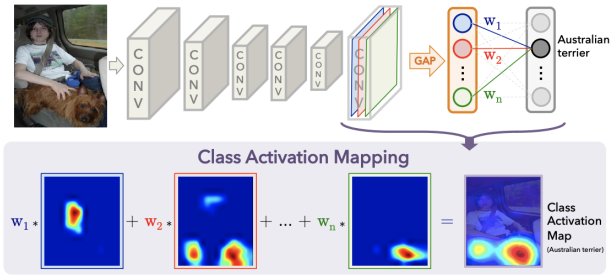


Figure 2: Class Activation Maps

1. **Average Drop**: Defined as the average percentage drop in confidence for the target class when using only the explanation map instead of the full image. For a single image it is calculated as $\left(\frac{max(0, y_c - o_c)}{y_c}\right) \times 100$, where $y_c$ is the output score for class $c$ when using the full image, and $o_c$ is the output score when using the explanation map. The value is then averaged over a set of images. It should be less for a better a good activation map.

2. **Coherency** : Similarity between the activation mapping of input image and that of only explainability map of the input image. The paper proposes pearson correlation between the images, but we can go for other similarity metrics like SSIM index. It should be more for a better activation map.

3. **Complexity**: Measured using the L1 norm as a proxy for the complexity of a CAM -

   $$Complexity(x) = ||CAM_c(x)||_1$$

   Complexity is less when the number of pixels highlighted by the attribution method is low, (narrowed area of focus). For a model with better explanations, this value would be low.

### 2.3. Exploring a Compression Technique - Knowledge Distillation

However there are many methods for CNN compression including Pruning, Low-rank Matrix Decomposition, Knowledge Distillation(KD), etc., this section explains about the KD method (Hinton et al., 2015) that we would be analyzing in the project.

Testing with KD for students begins with training a teacher model. This involves the standard process of using cross-entropy loss based on the actual labels. The initial step in knowledge distillation is to transfer knowledge from a

conventionally trained teacher model to transform the pre-softmax logits ($z_i$) computed for each class into probabilities ($q_i$). This transformation is done using the equation

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

Here, *zi* represents the pre-softmax logits for each class, *T* is the temperature parameter, and *qi* denotes the softened probability for class *i*. A higher temperature results in a 'softer' probability distribution across classes. With softened softmax scores, information from incorrect classes might become more evident for distillation. The knowledge obtained from training a teacher model with a regular softmax (i.e., $T = 1$) can be transferred to a student network through a new knowledge distillation loss ($L_{kd}$).

$$L_{kd}(W_{student}) = \alpha \cdot T2 \cdot CrossEntropy(Q_s, Q_t) +$$

$$(1 - \alpha) \cdot CrossEntropy(Q_s, y_{True})$$

This loss function, $L_{kd}(W_{student})$ , involves two components. The first part incorporates $\alpha \cdot T^2 \cdot CrossEntropy(Q_s, Q_t)$, where $Q_s$ and $Q_t$ represent the softened "targets" of the student and teacher, both using the same temperature $T$ ($> 1$). Another hyperparameter, $\alpha$ tunes the weighted average between the two loss components. The aim of this first component is to guide optimization toward similar softened softmax distributions for the student.

The second component in the expression, $(1 - \alpha) \cdot CrossEntropy(Q_s, y_{True})$, directs optimization toward approximating the ground truth labels as usual. $\alpha$ serves as a parameter where $\alpha = 1$ signifies using only distilled knowledge with 'unlabeled' data during student training. Adjusting alpha allows controlling the balance between utilizing the distilled knowledge and adhering to ground truth labels for student training.

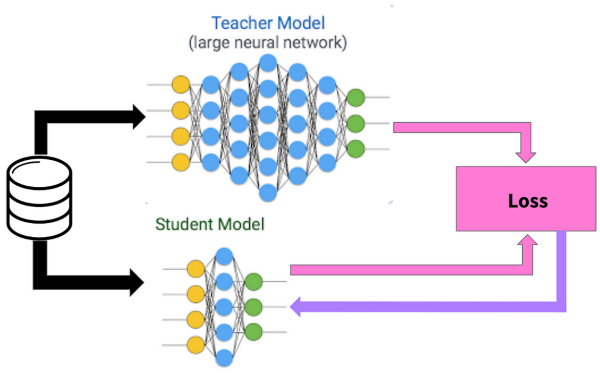


Figure 3: Knowledge Distillation

## 3. Methodology

We used the models trained on CIFAR-10 dataset for our analysis. The compression technique that we tend to analyze in the project is Knowledge Distillation. The teacher model considered for this is Wide Residual Network(WRN-28-10), which is a high computation model - a widened version of ResNet. It has a depth of 28 layers and widening factor of 10. The student model taken for analysis is Resnet-18 (figure - 4), which is simpler and more computation efficient than the teacher model since it has lesser depth of 18 layers. So the compression from the teacher to student model should be more efficient.

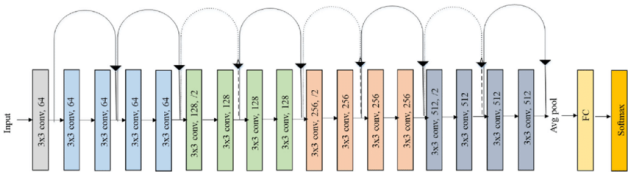

Figure 4: Student Model : ResNet-18

The student model underwent training using knowledge distillation from teacher model. The aim was to achieve comparable accuracies in the student model, aligning with the performance of the teacher model. The distilled student model showed an accuracy of 94%.

For gradient-based explainability map, we employed Class Activation Maps (CAM) on both the teacher and distilled student model. Next, we took random 1000 samples from the dataset and we calculated the metrics Average Drop Value, Complexity and Coherency(SSIM index) for the teacher and student models.

This comprehensive methodology ensures a rigorous analysis of the impact of knowledge distillation on gradient-based explainability in compressed CNNs.

## 4. Results

The class activation maps produced for some image inputs can be found in the figures 5 and 6 for both teacher and student models. The first row of images is the original image and the second row is the CAM of the image for the predicted class for that model.

Using these CAMs of the input images, the metrics proposed are calculated for both teacher and student model and are listed in the table 1.

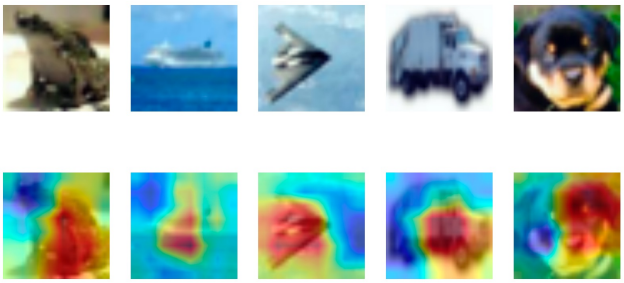

Figure 5: CAMs of Teacher Model : WRN-28-10

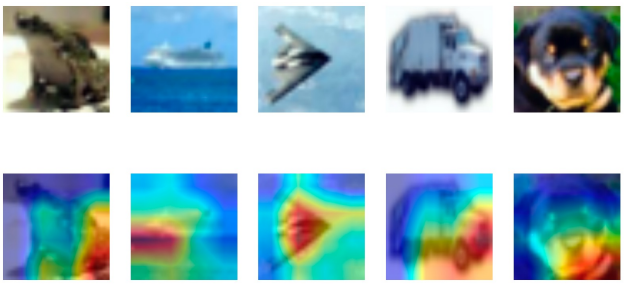

Figure 6: CAMs of Student Model : ResNet-18

| Model | WRN-28-10 | Distilled Resnet-18 |
|---|---|---|
| **Avg Drop Value** | 12% | **8%** |
| **Complexity** | 5.65 | **2.73** |
| **SSIM Index** | 0.58 | **0.67** |

Table 1: Resulting Metric Values

## 5. Discussion

### 5.1. Interpretation of Results

We observed and compared many activation maps for both the models. One thing that we noticed is that the area of focus of the map in the student model changes slightly as compared to the teacher model. However both the maps highlight some relevant features in the input image. But there are differences in the exact locations of the focus, and by looking, the teacher model explanations seem slightly more accurate.

The objective metrics however show interesting results. The average confidence drop for the student resnet model is 8% which is slightly less than that of the teacher model (12%), signifying that the explanations drawn by the student model are somewhat more confident in giving the right prediction, when only explanation maps are given as input. The complexity of the student model is also lesser than the teacher model which lets it highlight lesser, but relevant pixels in the explanation. Lastly, the coherency value(SSIM index) for the two CAMs is also more in the case of student model which ensures that its CAM contains more relevant features that explain a prediction and removes useless features in a more coherent way than the techer model. With these results, we can infer that model compression can result in better explainability if the student model is deep enough to capture relevant features. We used the state-of-the-art model ResNet-18 which may be the reason for the better explainability based metric values. With a more shallower model, it is intuitive that the feature capturing would not be that great.

### 5.2. Comparison with Existing Work

The paper (Weber et al., 2023) proposes that there exists a sweet spot for compression and explainability, while experimenting with pruning compression techniques. For slight compression factor of 2, the gradient-based explainability is found to increase a little bit, while for aggressive compression factors like 16 and 32, the explanation dips steeply. This is consistent with the results that we got, since we compressed the teacher model by Deep Knowledge Distillation by taking a deep model like ResNet-18. And for the aggressive compression, if we take a very shallow model of let's say 4-5 layers, we expect the explanation to be very inconsistent with the model prediction due to very high complexity of the teacher model.

### 5.3. Limitations

Since there is not an ideal explainability method for drawing explanations, there's not a right way to approach the problem of analyzing compression of CNNs. While the gradient-based methods like CAM, GradCAM, Guided Backpropagation, etc. are used in the context, they may not be fully depended on the model as proposed by (Adebayo et al., 2018). So, new and more efficient methods for model explanations are required to be researched and proposed. But for the scope of this project we went with CAMs for drawing explanations. Also, the training for Knowledge Distillation is very resource heavy, so we couldn't experiment much with other pairs of teacher and student models for ensuring that our results are consistent with those.

## 6. Conclusion

In this project, we explored the intricate relationship between CNN compression, specifically through knowledge distillation, and the consequent impact on gradient-based explanability. Our investigation focussed on the complex teacher model WRN-28-10, and the simpler student model Resnet-18, training the latter through knowledge distillation to align with the performance of the former.

We used the metrics proposed in the paper (Poppi et al., 2021) average drop, coherency, and complexity to infer the changes in the distilled model explainability, offering

a comprehensive evaluation. Notably, as efficiency gains were achieved through compression, interpretability metrics exhibited subtle changes, possibly implying a more explainable compressed model.

## 7. Future Work

This study majorly focussed on classification tasks using CNN, which can be extended to other deep neural network architectures such as Inception, RNNs or Vision Transformers. A similar analysis could be made for other computer vision tasks like segmentation and object detection. Other than CAM, novel gradient-based interpretability methods like GradCAM, GradCAM++, etc. can be used to constuct similar metrics and analysis. Looking ahead, the study lays a foundation for future investigations, prompting consideration of refined compression techniques and novel interpretability methods. As the landscape of AI continues to evolve, the quest for an optimal equilibrium between computational efficiency and interpretable decision-making remains an open challenge, inviting further exploration and innovation.

## 8. Links to resources

1. Dataset : CIFAR10
2. Code : https://github.com/vaibhavchhabra25/XML

## References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. Sanity checks for saliency maps. *CoRR*, abs/1810.03292, 2018. URL http://arxiv.org/abs/1810.03292.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015.

Poppi, S., Cornia, M., Baraldi, L., and Cucchiara, R. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. *CoRR*, abs/2104.10252, 2021. URL https://arxiv.org/abs/2104.10252.

Weber, D., Merkle, F. T., Schöttle, P., and Schlögl, S. Less is more: The influence of pruning on the explainability of cnns. *ArXiv*, abs/2302.08878, 2023. URL https://api.semanticscholar.org/CorpusID:257019751.