DOWNSAMPLING BLOCKS AND RESIDUAL CONNECTIONS PROPAGATE MEMORIZATION ENTANGLED WITH GENERALIZATION IN RESNETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Residual networks (ResNets) are among the most widely used deep learning models due to their ability to train very deep neural networks through a stable gradient flow enabled by residual/skip connections. However, no prior work has studied how different model pathways propagate memorization and generalization within the ResNet architecture. Hence, in this work, we investigate which pathways in ResNets are responsible for propagating memorization and generalization. We find that downsampling blocks and residual connections propagate memorization along with generalization, supported by activation-level analysis. We conduct a block-wise Fisher information analysis to elucidate the dominant role played by downsampling blocks in propagating memorization. We also provide a weight-level analysis to show that weights contributing to memorization and generalization are entangled and primarily concentrated in downsampling blocks. We validate our findings across three interpretability strategies, six architectures, and six datasets. Together, our results provide a clear map of how memorization and generalization propagate in the ResNet architecture, which has never been articulated in the literature.

1 INTRODUCTION

Residual Networks (ResNets) introduced by (He et al., 2016), are one of the most widely adopted architectures in deep learning history. They were devised to mitigate the vanishing gradient problem by adding residual/skip connections (a.k.a. identity mappings). These connections facilitate smoother gradient flow, faster convergence, and more effective feature reuse across layers (Hardt & Ma, 2016; Liu et al., 2019; Philipp et al., 2018). Overall, their modularity, scalability, and robustness make them foundational models throughout modern deep learning.

Despite ResNets having shown to exhibit remarkable generalization ability on real world data, they also have a tendency to memorize training data, especially in over-parameterized regimes, a term commonly known as *label memorization* (Feldman, 2020; Feldman & Zhang, 2020). As described in prior work, memorization refers to the phenomenon where the model memorizes certain training data points without learning generic patterns, and hence does not generalize well on the test set (Arpit et al., 2017; Singhal & Kim, 2025). Recent studies have shown that neural networks tend to learn easier samples and memorize hard samples (Baldock et al., 2021; Stephenson et al., 2021; Zhou & Wu, 2023). Easy samples generally straightforwardly represent the ground truth class, while harder samples tend to be visually atypical, confusing, and/or at times mislabeled (Agarwal et al., 2022). Besides understanding which samples tend to be memorized, other studies focus on isolating the components responsible for memorization in ResNets. Prior work has suggested that memorization tends to emerge in the later layers of deep networks (Baldock et al., 2021; Stephenson et al., 2021), while more recent findings argue that, it is instead dispersed across multiple layers (Maini et al., 2023). However, these analyses focus only on the parameterized layers and overlook the residual/skip connections, which also carry signals that may encode either or both memorization and generalization.

In this work, we discover that downsampling blocks and residual connections propagate memorization along with generalization. Specifically, residual connections across all blocks and main branch in downsampling blocks propagate memorization & generalization, supporting these distinctive impacts

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

via an activation-level analysis. We also analyze the trace of Fisher Information Matrix to understand the role of downsampling blocks in memorization. Lastly, we also present a weight-level analysis, elucidating that the weights contributing to memorization and generalization are entangled and primarily concentrated in downsampling blocks. We rigorously validate our claims across three interpretability strategies, six models, and six datasets. Please note that the aim of this study is to provide novel insights into the pathways responsible for memorization and generalization in ResNets, rather than to offer a specific solution. In summary, the novel insights and core contributions of our paper are as follows:

- We observe that downsampling blocks and residual connections propagate memorization along with generalization, supporting it via activation-level analysis.
- We show that downsampling blocks exhibit high Fisher information matrix trace, to understand their dominant role in memorization against non-downsampling blocks.
- We find that the weights contributing to memorization and generalization are entangled and primarily concentrated in downsampling blocks.

2 Related Works

Residual Networks (ResNets): Residual Networks (ResNets) (He et al., 2016) introduced *residual connections* (a.k.a. skip connections or identity mappings), a special case of highway networks (Srivastava et al., 2015). These connections were designed to mitigate the vanishing gradient problem, a primary cause of difficulty in model optimization (Bengio et al., 1994; Pascanu et al., 2013; Schoenholz et al., 2016). This shortcut mechanism provides an alternative path for information to flow in addition to the main branch stream, while improving gradient flow and training stability (Philipp et al., 2018; Balduzzi et al., 2017). Several studies have shown that residual connections improve convergence, help avoid spurious local optima (Liu et al., 2019; Hardt & Ma, 2016), and enhance generalization compared to standard deep feed-forward networks (Huang et al., 2020; Szegedy et al., 2017). Furthermore, (Veit et al., 2016; Abdi & Nahavandi, 2016) demonstrated that ResNets can be interpreted as an ensemble of many short paths. The closest related work by (Veit et al., 2016) measures the influence of removing the main branch on test accuracy while retaining the residual connections. However, it neither examined the impact of the main branch on memorization nor articulated how residual connections propagate memorization and generalization. ***With the best of our knowledge, no prior work has systematically investigated how model pathways in ResNets–residual connections and main branch, influence memorization and generalization.***

Memorization and Generalization: Neural networks demonstrate a remarkable ability to extract rich, generic patterns from training data and to generalize effectively to unseen test samples (Arpit et al., 2017; Shah et al., 2020). However, due to their over-parameterized nature, they are also prone to memorizing rare, complex, or mislabeled samples, which can impair generalization performance—a phenomenon commonly referred to as *label memorization* (Feldman, 2020; Feldman & Zhang, 2020). Prior works have focused on identifying the types of samples that are most susceptible to memorization, such as those with high prediction depth (Baldock et al., 2021), high loss curvature (Garg et al., 2023; Ravikumar et al., 2024), high consistency score (Jiang et al., 2020), mislabeled/rare (Maini et al., 2023; Singhal & Kim, 2025), or easily forgotten (Toneva et al., 2018). Beyond characterizing memorized samples, several studies have attempted to localize memorization within neural networks. While some works (Baldock et al., 2021; Stephenson et al., 2021) argue that memorization primarily occurs in the later layers of the network, others (Maini et al., 2023) suggest that it is broadly spread across layers. These conflicting findings raise a fundamental question: ***Which pathways and blocks are responsible for propagating memorization within ResNet?***

3 Preliminaries

3.1 Structure of ResNets

ResNet consists of four stages, and each stage consists of multiple blocks. In each stage, the first block is a downsampling block, and the remaining blocks are non-downsampling blocks. The two types of blocks differ in how they transform feature dimensions. For a non-downsampling block, the skip connection is an identity mapping, with the output as follows:

$$y = x + F(x), \tag{1}$$

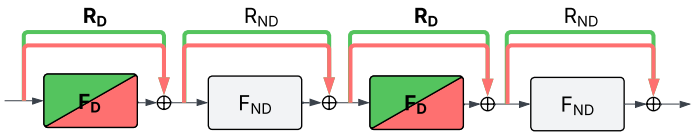


Figure 1: **Illustrating the flow of memorization and generalization across residual connections and main branch in ResNets.** Downsampling blocks and residual connections propagate memorization along with generalization. Specifically, both Residual connections (R_D) and main branch (F_D) in downsampling blocks carry them, while residual connections (R_{ND}) in non-downsampling blocks are the primary carriers. Generalization is marked in green, while memorization is in red.

where $x \in \mathbb{R}^{H\times W\times C}$, $y \in \mathbb{R}^{H\times W\times C}$, and $F(x)$ is the learned mapping consisting of convolutional & batch-norm layers. Here, H and W denote the spatial dimensions and C denotes channel count.

On the other hand, a downsampling block decreases spatial dimensions while increasing the number of channels (except for stage 1, where it only increases channels, but there spatial downsampling has already been performed by the preceding layer through max-pooling and convolution with a stride of 2). The skip connection uses a projection W_s to match dimensions, and the block computes:

$$y = W_s x + F_s(x), \tag{2}$$

where, $x \in \mathbb{R}^{H\times W\times C}, y \in \mathbb{R}^{H'\times W'\times C'}, H' = \frac{H}{s}$, $W' = \frac{W}{s}$, $C' = kC$. Here, W_s is a 1×1 convolution used for matching the output size, and $F_s(x)$ denotes the main branch containing a strided convolution operation with stride $s > 1$, and channel expansion via $k > 1$.

3.2 Understanding Memorization and Generalization

Deep neural networks are supposed to understand generic features and patterns from the training data and generalize well to the test set, which is known as ***generalization***. However, they can also memorize specific examples (generally ambiguous, complex, atypical, and/or mislabeled) instead of learning underlying generic patterns, a behavior referred to as ***label memorization***. In this study, we investigate memorization by injecting noisy labels into the training set–a widely adopted method to study memorization (Maini et al., 2023; Singhal & Kim, 2025). Specifically, we alter the label of a portion of training samples by assigning them a randomly chosen class that differs from the true class. To ensure that these mislabeled samples are memorized by the network, we train the model until it attains perfect ($\approx 100\%$) training accuracy. To rigorously validate our claims, we vary the proportion of mislabeled samples as 2%, 5%, and 10% of the train set (Please note that, unless specified, the proportion of mislabeled samples is 2%). Furthermore, to ensure robustness in our claims, we evaluate all results for three random seeds.

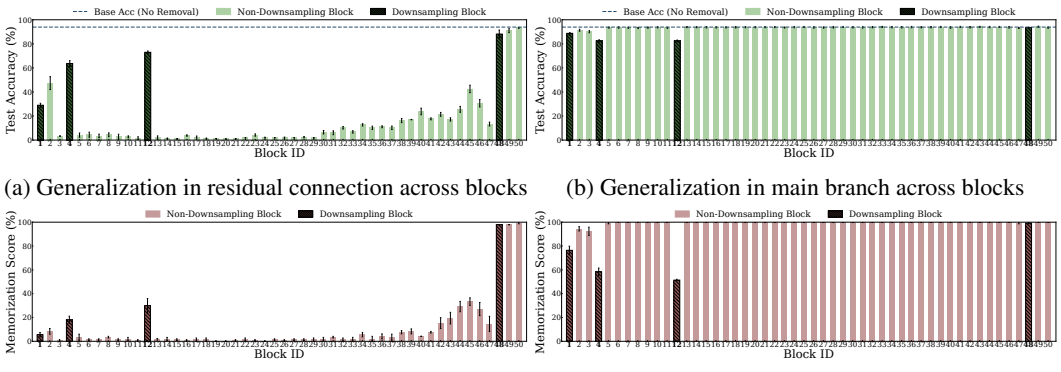
3.3 Investigating the Impact of Residual Connections and the Main Branch on Memorization and Generalization

To identify which pathways in the network contribute to memorization and generalization in ResNets, we utilize three commonly adopted ablation approaches: zero ablation (Veit et al., 2016; Bhojanapalli et al., 2021), mean ablation (Tigges et al., 2023; McDougall et al., 2023; Ghorbani & Zou, 2020), and optimal ablation (Li & Janson, 2024), to show that our findings are not tied to a specific ablation strategy. We provide a detailed discussion about the three strategies in Appendix B.

3.4 Metrics to Measure Memorization and Generalization

To evaluate the impact of each model pathway — *residual connection* and *main branch*, on memorization and generalization in ResNet, we use the metrics proposed in Maini et al. (2023); Singhal & Kim (2025), memorization score and test accuracy, defined as follows:

Memorization Score (%): A degree to overfit on the noisy labels in the training set, i.e., memorize them, rather than learning any generic patterns from them ($\frac{\#\text{Noisy label samples memorized}}{\#\text{Total noisy label samples}} \times 100$). A higher score indicates the model is "memorizing" the noisy labels.



(a) Generalization in residual connection across blocks

(b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks

(d) Memorization in main branch across blocks

Figure 2: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (Caltech101, ResNet-152, Optimal Ablation).

Test Accuracy (%): Model's ability to learn meaningful generic patterns during training and then generalize them to the unseen test set, an indicator of the model's generalization ability ($\frac{\#\text{Correct predictions on test set}}{\#\text{Total test set samples}} \times 100$).

3.5 Datasets and Models

We validate our claims using 6 models and 6 datasets, with more details in Appendix A:

Datasets: CIFAR10 (Krizhevsky et al., 2009), Places365 (Zhou et al., 2017), TinyImageNet (Le et al., 2015), Oxford-IIIT Pet (Parkhi et al., 2012), Oxford Flower-102 (Nilsback & Zisserman, 2008), and Caltech101 (Fei-Fei et al., 2004).

Models: ResNet152, ResNet101, ResNet50, ResNet34 (He et al., 2016), WideResNet101 (Zagoruyko & Komodakis, 2016), and ResNeXt101 (Xie et al., 2017).

4 Downsampling Blocks and Residual Connections Propagate Memorization along with Generalization

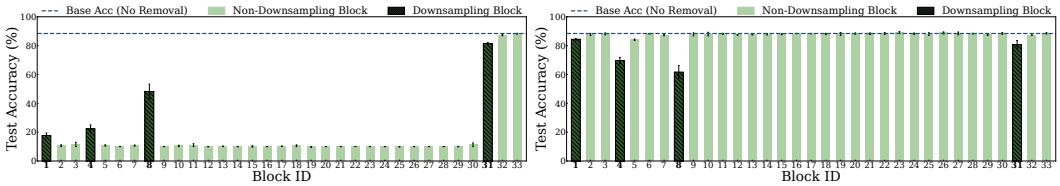
In this section, we study how different model pathways–residual (skip) connections and main branch propagate memorization and generalization across the ResNet architecture, by following the investigation setup discussed in Section 3.3. We divide our analysis into downsampling and non-downsampling blocks since ResNet consists of these two types of blocks as discussed in Section 3.1.

4.1 Memorization and Generalization in Downsampling Blocks

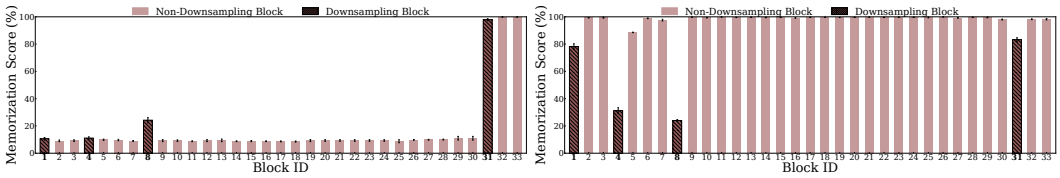
From Fig. 2, we can observe that when we ablate the residual connections and the main branch pathways independently in the downsampling blocks, both memorization and test accuracy decrease. This essentially depicts that downsampling blocks as a whole propagate memorization and generalization via both residual connections and main branch. Consistent results are observed for other ablation strategies, remaining architectures–ResNet101, ResNet152, ResNet50, ResNet34, WideResNet101, and ResNeXt101, and datasets–Places365, Oxford Flower-102, Caltech101, Oxford-IIIT Pet, and TinyImagenet, as presented in Appendix C.3.1, C.3.2, and C.1.

4.2 Memorization and Generalization in Non-Downsampling Blocks

Having established that downsampling blocks encode both memorization and generalization, we now examine the internal pathways of non-downsampling blocks. As shown in Figs. 2d & 2b, ablating the main branch in non-downsampling blocks has a negligible impact on both memorization and



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks



(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 3: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. On the other hand, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Places365, ResNet-101, 5% Label Noise, Zero Ablation).

generalization, depicting their minimal role in propagating them. Since we know from Section 4.1 that the preceding downsampling block is passing both memorization and generalization information forward, if the main branch propagates them negligibly, then the only remaining pathway available to carry that information is the “residual connection". This role of the residual connections is further confirmed by the sharp drop in both memorization and test accuracy when residual connections are ablated, as seen in Figs. 2c & 2a. Note that the drop in both memorization and generalization is much more severe than ablating residual connections in downsampling blocks, because main branch in non-downsampling blocks has almost negligible impact on both memorization and generalization. Therefore, removing residual connections in these blocks essentially removes the dominant carrier of memorization and generalization. Apart from this, we also observe another interesting phenomenon where ablating residual connections in the last two non-downsampling blocks has a negligible impact on both memorization and generalization, unlike the residual connections in other non-downsampling blocks. Consistent results are observed for other ablation strategies, remaining architectures–ResNet101, ResNet152, ResNet50, ResNet34, WideResNet101, and ResNeXt101, and datasets–Places365, Oxford Flower-102, Caltech101, Oxford-IIIT Pet, and TinyImagenet, as presented in Appendix C.3.1, C.3.2, and C.1. Note that while the main branch’s contributions to generalization across downsampling and non-downsampling blocks align with (Veit et al., 2016), the work did not explore its role in memorization and never articulated how residual connections propagate both phenomena across the network.

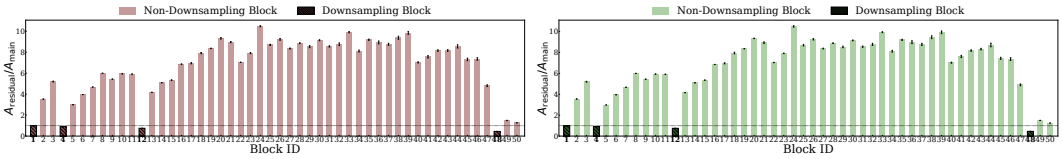
4.3 Consistent Trends across Higher Label Noise Ratios

To further reinforce the robustness of our findings, we evaluate all models (ResNet152, ResNet101, ResNet50, ResNet34, WideResNet101, and ResNeXt101) and datasets (TinyImageNet, CIFAR10, Places365, Oxford-IIIT Pet, Oxford Flower-102, and Caltech101) under higher label noise ratios of 5%, and 10%, apart from 2%, and present them in Fig. 3 & Appendix C.2.

Pathway	**Residual**		**Main**	
Block Type	**Mem**	**Gen**	**Mem**	**Gen**
Non-Down	✓	✓	✗	✗
Down	✓	✓	✓	✓

Table 1: Summarizing block-wise (**Down**sampling, **Non-Down**sampling) and Path-wise (**Residual** Connection and **Main** Branch) effects on memorization (**Mem**) and generalization **(Gen)**.

From the results we can again confirm that, downsampling blocks and residual connections propagate memorization along with generalization, with main branch in non-downsampling blocks having a negligible impact on both phenomena. These distinctive roles of downsampling vs. non-downsampling blocks and residual connections vs. main branch on memorization and generalization, are summarized in Table 1 and Fig. 1.



(a) Activation Analysis for Memorization (b) Activation Analysis for Generalization

Figure 4: **Activation-level analysis to understand how differently residual connections and main branch propagate memorization and generalization.** We observe that in downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ depicting that both residual connections and main branch are propagating memorization and generalization. Conversely, in most non-downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \gg 1$, showing that residual (skip) connections are the primary carriers. Furthermore, in the last two non-downsampling blocks (Block IDs 49, 50) $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ confirming the observation made in Fig. 2 that skip connections and main branch in those blocks have a similar minimal impact on memorization and generalization (CIFAR10-ResNet152).

5 Activation-Level Analysis to Understand How Residual Connections and Main Branch Propagate Memorization and Generalization

In Section 4, we established that downsampling blocks and residual connections drive the propagation of both memorization and generalization, with the main branch in non-downsampling blocks having a negligible impact. To better understand this behavior, we analyze the L_2 norm of the activations for both pathways (residual connections and main branch) across all residual blocks.

Because downsampling blocks use both the residual connections and the main branch to propagate memorization and generalization, we expect their mean activation levels to be comparable for both phenomena. Conversely, in non-downsampling blocks—where residual connections serve as the primary carriers for memorization and generalization—residual connections activations are expected to exceed main branch activations. To test this, we compute the ratio of the L_2 norm of the activations (averaged across mislabeled samples for memorization and test set samples for generalization) corresponding to residual connections and main branch ($\frac{A_{\text{residual}}}{A_{\text{main}}}$) across all blocks.

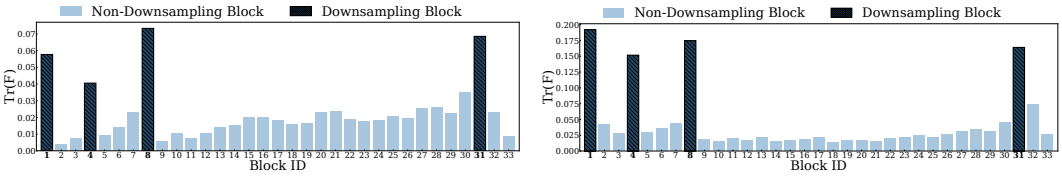
From Fig. 4, we can observe that downsampling blocks yield $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ for both memorization and generalization, confirming that both pathways contribute towards memorization and generalization in downsampling blocks. In contrast, most non-downsampling blocks exhibit $\frac{A_{\text{residual}}}{A_{\text{main}}} \gg 1$, validating the dominance of residual connections in propagating memorization and generalization compared to the main branch. One minor note is that, in the final two non-downsampling blocks, the ratio drops to $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$. This aligns with our observations in Fig. 2, where both pathways exert a similarly minimal impact on memorization and generalization in the final two non-downsampling blocks. Consistent results are observed for remaining models–ResNet101, ResNet50, ResNet34, WideResNet101, & ResNext101 and datasets–Places365, Oxford Flower-102, Oxford-IIIT Pet, & Caltech101, presented in Appendix C.5.

6 Understanding Downsampling Blocks' Role to Propagate Memorization Via Fisher Information Matrix Trace

In this section, we aim to better understand the prominent impact of downsampling blocks on memorization, as observed in Section 4.

6.1 Downsampling Blocks Exhibit Higher Fisher Information Trace than Non-Downsampling Blocks

Prior work (Jastrzebski et al., 2021) studies the cause of poor generalization in CNNs through Fisher Information Matrix (FIM), which approximates the local curvature of the loss landscape (Thomas et al., 2020; Martens, 2020). They showed that a high trace of the FIM, $\mathrm{Tr}(\mathrm{F})$, computed



(a) Trace of Fisher Information Matrix, $\mathrm{Tr}(\mathrm{F})$, across blocks (Places365-ResNet101)

(b) Trace of Fisher Information Matrix, $\mathrm{Tr}(\mathrm{F})$, across blocks (Oxford Flower-102-ResNeXt101)

Figure 5: **Downsampling blocks exhibit higher $\mathrm{Tr}(\mathbf{F})$ than non-downsampling blocks.** We compute the block-wise Fisher information trace $\mathrm{Tr}(\mathrm{F})$ for each block in ResNet101 and ResNeXt101. In both cases, downsampling blocks consistently exhibit much higher $\mathrm{Tr}(\mathrm{F})$ than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization. Consistent trends are observed for remaining models and datasets, as shown in Appendix C.4.

throughout the network, is indicative of memorization and poor generalization, as larger values of $\mathrm{Tr}(\mathrm{F})$ correspond to sharper loss curvature and have been empirically linked to memorization (Garg et al., 2023; Ravikumar et al., 2024; Jeon et al., 2024). Mathematically, $\mathrm{Tr}(\mathrm{F})$ is denoted as:

$$\mathrm{Tr}(\mathrm{F}) = \mathbb{E}_{(x,y)\sim\mathcal{D},\ \hat{y}\sim p_\theta(y|x)}\left[\left\|\frac{\partial \ell(y,\hat{y})}{\partial \theta}\right\|_2^2\right] \tag{3}$$

where θ are the model parameters, (x, y) is sampled from the dataset $\mathcal{D}$, $\hat{y}$ is the prediction vector produced by the model for input x, and $\ell(y, \hat{y})$ is the loss function.

Motivated by this observation, we ask which blocks are responsible for the elevated Fisher information within the network. This question is particularly relevant given our findings in Section 4, where we identify that downsampling blocks propagate memorization. Please note that whenever we refer to downsampling/non-downsampling block w.r.t fisher information matrix trace, we are referring to the main branch within each of them. We do not analyze the residual connection part in both blocks because in non-downsampling blocks, the residual connections do not consist of any weight parameters, and hence it is not feasible to compute the Fisher information matrix trace for them. All reported trace values are normalized by the number of parameters in the corresponding block for fair comparisons among blocks. Also, we use Fisher information matrix trace as a diagnostic, correlational tool (not causal) to understand the primary role of downsampling blocks in memorization.

Through this analysis, we observe from Figs. 5a & 5b that downsampling blocks consistently exhibit higher $\mathrm{Tr}(\mathrm{F})$ values compared to non-downsampling blocks. This indicates that downsampling blocks contribute disproportionately to the overall $\mathrm{Tr}(\mathrm{F})$ of the network, which, according to (Jastrzebski et al., 2021), is associated with memorization and poor generalization. Hence, this correlation provides further support for our observation that downsampling blocks play a prominent role in memorization in ResNets. Consistent results are observed for remaining models–ResNet101, ResNet152, ResNet50, ResNet34, WideResNet101, and ResNeXt101, and datasets–Places365, Oxford Flower-102, Caltech101, and Oxford-IIIT Pet, as discussed in Appendix C.4.

7 Weights Contributing to Memorization and Generalization are Entangled and Primarily Concentrated in Downsampling Blocks

In Section 4 we observe that the downsampling blocks and residual connections propagate memorization along with generalization by performing pathway ablation. In this section, we examine, at finer granularity, whether weights contributing to memorization and generalization are primarily concentrated in the downsampling blocks. To carry out this analysis, we utilize one of the most fundamental weight importance estimation methods based on Taylor First Order approximation (Molchanov et al., 2019). This method estimates the global importance of a weight (w_i) for a given task (depicted by score S_i), based on the magnitude of the weight and the corresponding gradient. It is written as follows:

$$S_i = \sum_{d\in D} \left| g_{i,d}\, w_i \right|, \tag{4}$$

where D represents the representative dataset for the task, d denotes a sample within D, and $g_{i,d}$ represents the loss gradient w.r.t w_i corresponding to the sample d. We compute this score for all weights within the residual blocks for datasets corresponding to memorization (mislabeled samples) and generalization (held out calibration validation set samples). We specifically focus on the weights in the convolutional layers.

From Fig. 6, we can clearly observe that the top-K=0.1% of memorization and generalization weights are primarily concentrated in the downsampling blocks, reinforcing our earlier finding that downsampling blocks play a central role in propagating both memorization and generalization. Consistent trends hold across all K ratios (0.0001%, 0.001%, 0.01%, and 0.1%), architectures–ResNet-152, ResNet-101, ResNet-34, ResNet-50, WideResNet101, and ResNeXt101, and datasets–CIFAR-10, Places365, Oxford Flowers-102, Caltech101, and Oxford-IIIT Pet, presented in Appendix C.6.1.

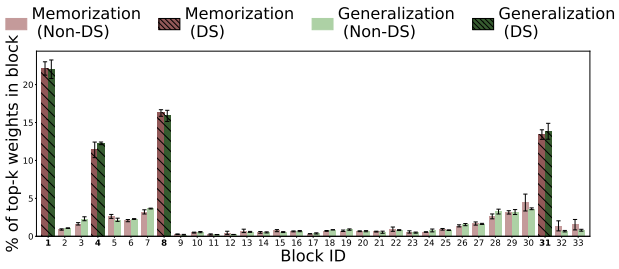


Figure 6: Memorization and Generalization weights are primarily concentrated in downsampling blocks (K=0.1%, Places365-ResNet101, DS: Downsampling, NDS: Non-Downsampling)

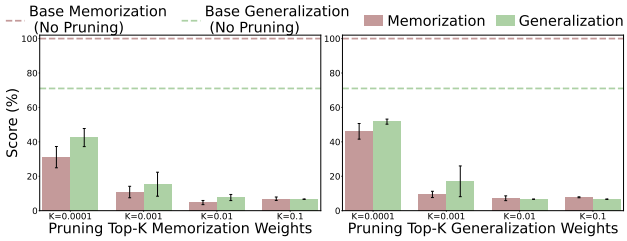


Figure 7: Weights contributing to Memorization and Generalization are entangled (Places365-ResNet101)

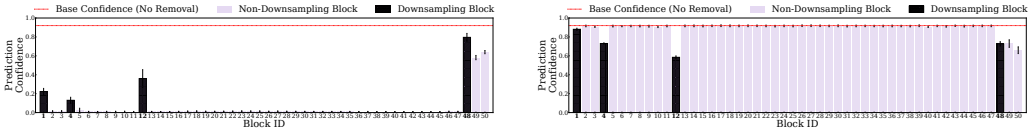
The aligning trends of memorization and generalization in the result of Fig. 6 hint that the two are entangled. To precisely examine it, we repeat this analysis across the same four top-K ratios (0.0001%, 0.001%, 0.01%, and 0.1%), by independently pruning the top-K memorization-ranked weights and the top-K generalization-ranked weights. Fig. 7 shows that pruning either set causes a substantial drop in *both* memorization and generalization relative to the unpruned baseline. Moreover, we observe substantial overlap between Top-K memorization and generalization weights (see Fig. 88), providing further evidence that memorization and generalization weights are entangled. To also show that these effects are not simply a consequence of pruning an arbitrary subset of weights, we show a reference of pruning randomly selected weights (see Fig. 89), where we see almost no impact on memorization and generalization.

Taken together, our analysis clearly indicates that memorization and generalization are entangled at the level of weights, where the weights responsible for one are not cleanly separable from the weights responsible for the other. Consistent trends are also observed for other models–ResNet152, ResNet101, ResNet34, ResNet50, WideResNet101, and ResNeXt101, and datasets–CIFAR10, Places365, Oxford Flower-102, Caltech101, and Oxford-IIIT Pet, as presented in Appendix C.6.2.

8 Prediction Confidence during Test-Time

In Section 4, we showed that downsampling and non-downsampling blocks play fundamentally different roles in propagating memorization and generalization. Building on these observations, we now analyze how the model's **prediction confidence** on the test set–probability of predicting the expected, correct class label during testing–changes when ablating either residual connections or main branch across different blocks. This is particularly informative because sometimes the test accuracy can be negligibly affected even when confidence in predicting the correct class is low. Formally, we define it as,

$$\text{Prediction Confidence} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x,y)\in\mathcal{D}_{\text{test}}} p_\theta(y \mid x) \tag{5}$$



(a) Impact of Residual Connections on Prediction Confidence across blocks (Caltech101-ResNet152)

(b) Impact of Main Branch on Prediction Confidence across blocks (Caltech101-ResNet152)

Figure 8: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 49, 50) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. Consistent trends observed for other architectures and datasets in Appendix C.7. (Results are averaged across the whole test set for 3 seeds, Zero Ablation.)

where $p_\theta(y \mid x)$ denotes the softmax probability assigned by the model to the correct, expected class label y for the input x in test set, $\mathcal{D}_{\text{test}}$.

Residual Connections Analysis: From Fig. 8a, we observe that removing residual connections from most *non-downsampling* blocks causes a substantial drop in prediction confidence. This is consistent with our findings in Section 4, where ablating residual connections in these blocks led to a significant decrease in test accuracy (Fig. 2a). In contrast, removing residual connections from *downsampling* blocks results in a comparatively smaller reduction in prediction confidence. This behavior also aligns with Fig. 2a, where ablating residual connections in downsampling blocks led to a less pronounced degradation in generalization, primarily because at downsampling blocks, generalization is split into main branch and residual connections.

Main Branch Analysis: From Fig. 8b, we observe that removing the main branch from non-downsampling blocks has a negligible effect on prediction confidence. This is consistent with our earlier observation that the main branch in non-downsampling blocks contributes minimally to generalization (see Fig. 2b). In contrast, ablating the main branch in downsampling blocks leads to a noticeable reduction in prediction confidence, indicating that the main branch in these blocks plays a role to model's generalization, as also seen in Fig. 2b.

Confidence Boosting via Last Two Non-Downsampling Blocks: Previously, we observed in Section 4.2, that residual connections and main branch in the last two non-downsampling blocks do not have significant impact on memorization and generalization. However, interestingly, in Figs. 8a, 8b, we observe a slight decrease in prediction confidence when ablating residual connection and main branch in the last two non-downsampling blocks (Block IDs 49, 50 for ResNet152), even though the test accuracy remains largely unchanged (see Figs. 2a, 2b). This suggests that, while the residual connections and main branch in these two blocks do not significantly impact the test accuracy, they are important for the model's ***confidence boosting*** towards the correct, expected class label. Consistent trends are observed for all models–ResNet152, ResNet101, ResNet50, ResNet34, WideResNet101, and ResNeXt101, and datasets–CIFAR10, Oxford Flower-102, Places365, Caltech101, and Oxford-IIIT Pet, as presented in Appendix C.7.

9 Conclusion

In this work, we elucidate how memorization and generalization propagate across distinct pathways (residual connections vs. main branch) and block types (downsampling vs. non-downsampling) within ResNet architecture. Crucially, we discover that downsampling blocks and residual connections propagate memorization along with generalization, supported by activation-level analysis. We show that downsampling blocks exhibit a higher Fisher information matrix trace, linking sharper loss curvature to memorization. Lastly, we also find that the weights contributing to memorization and generalization are entangled and primarily concentrated in downsampling blocks. Together, these findings offer a mechanistic understanding of how memorization and generalization are propagated within ResNets. We discuss limitations in Appendix D.

AI USE STATEMENT

We have used AI tools for coding purposes and for polishing the text. We did not use AI tools for ideation or experimental setups. The authors have reviewed all AI-assisted content and take full responsibility for the content of the work.

REPRODUCIBILITY STATEMENT

For training our models, we utilized NVIDIA's 6 RTX-4090 GPUs and provide the code as a supplementary file.

REFERENCES

Masoud Abdi and Saeid Nahavandi. Multi-residual networks: Improving the speed and accuracy of residual networks. *arXiv preprint arXiv:1609.05672*, 2016.

Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10368–10378, 2022.

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*, pp. 233–242. PMLR, 2017.

Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. *Advances in Neural Information Processing Systems*, 34:10876–10889, 2021.

David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International conference on machine learning*, pp. 342–350. PMLR, 2017.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10231–10241, 2021.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

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.

Isha Garg, Deepak Ravikumar, and Kaushik Roy. Memorization through the lens of curvature of loss function around samples. *arXiv preprint arXiv:2307.05831*, 2023.

Amirata Ghorbani and James Y Zou. Neuron shapley: Discovering the responsible neurons. *Advances in neural information processing systems*, 33:5922–5932, 2020.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

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*, pp. 770–778, 2016.

Kaixuan Huang, Yuqing Wang, Molei Tao, and Tuo Zhao. Why do deep residual networks generalize better than deep feedforward networks?—a neural tangent kernel perspective. *Advances in neural information processing systems*, 33:2698–2709, 2020.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pp. 4772–4784. PMLR, 2021.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding memorization in generative models via sharpness in probability landscapes. *arXiv preprint arXiv:2412.04140*, 2024.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. Characterizing structural regularities of labeled data in overparameterized models. *arXiv preprint arXiv:2002.03206*, 2020.

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

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. *Advances in Neural Information Processing Systems*, 37:109233–109282, 2024.

Tianyi Liu, Minshuo Chen, Mo Zhou, Simon S Du, Enlu Zhou, and Tuo Zhao. Towards understanding the importance of shortcut connections in residual networks. *Advances in neural information processing systems*, 32, 2019.

Pratyush Maini, Michael C Mozer, Hanie Sedghi, Zachary C Lipton, J Zico Kolter, and Chiyuan Zhang. Can neural network memorization be localized? In *International Conference on Machine Learning*, 2023.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*, 2023.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11256–11264. IEEE, 2019.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. Pmlr, 2013.

George Philipp, Dawn Song, and Jaime G Carbonell. Gradients explode-deep networks are shallow-resnet explained. 2018.

Deepak Ravikumar, Efstathia Soufleri, Abolfazl Hashemi, and Kaushik Roy. Unveiling privacy, memorization, and input curvature links. *arXiv preprint arXiv:2402.18726*, 2024.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

Rishi Singhal and Jung-Eun Kim. Impact of layer norm on memorization and generalization in transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

Cory Stephenson, Suchismita Padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and SueYeon Chung. On the geometry of generalization and memorization in deep neural networks. *arXiv preprint arXiv:2105.14602*, 2021.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3503–3513. PMLR, 2020.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models, 2023. *URL https://arxiv. org/abs/2310.15154*, 2023.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Xiaoling Zhou and Ou Wu. Which samples should be learned first: Easy or hard? *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):15264–15278, 2023.

APPENDIX

A TRAINING DETAILS

In this section, we discuss the details of the datasets and models used in this paper, along with the experimental setup details associated with them.

A.1 DATASETS USED

In this paper, we considered 6 different datasets and discussed them in detail below:

CIFAR10: proposed in (Krizhevsky et al., 2009), consists of everyday life images spread across 10 classes, with a total of 50,000 images. We consider 40,000 images for training, 5,000 images for validation and test set respectively.

Places365: A large-scale scene recognition dataset originally introduced by (Zhou et al., 2017). In this paper, we use the Places365-Mini version available on Huggingface[1]. It consists of a total of 8,000 samples (7,500 in train set and 500 in test set, provided by Huggingface), spread across 10 classes.

Oxford-IIIT Pet: A fine-grained image classification dataset introduced by (Parkhi et al., 2012), containing 37 categories of pet breeds (25 dog breeds and 12 cat breeds), with roughly 3,600 images in train and test sets respectively, available at Huggingface[2].

TinyImagenet: A subset of ImageNet introduced by (Le et al., 2015), comprising 200 classes with 500 training images and 50 validation images per class, available via Hugging Face[3].

Oxford Flower 102: A flower classification dataset introduced by (Nilsback & Zisserman, 2008), consisting of around 8,000 images with 102 flower categories. The standard train-test split is provided by Huggingface[4].

Caltech101: An object recognition dataset introduced by (Fei-Fei et al., 2004), containing 101 object categories, with approximately 9,000 images in total. We use the standard train-test split available in Huggingface[5].

A.2 MODELS USED

In this paper, we specifically focus on the ResNet architecture proposed in (He et al., 2016), and therefore we perform our analysis across standard ResNet models of varying depths, publicly available on Pytorch. To reinforce our evaluations, we also experiment with commonly known ResNet variants–WideResNet (Zagoruyko & Komodakis, 2016) and ResNeXt (Xie et al., 2017). Across all models, there are 4 blocks consisting of downsampling operation. For the conv_3, conv_4, and conv_5 stages, the first block performs both spatial downsampling and channel expansion, except for the first block in the conv_2 stage, which only increases the number of channels, as spatial downsampling has already been performed by the preceding stride-2 convolution and max-pooling layers. Except for these traits, each architecture has the following distinctions:

ResNet34: A 34-layer residual network composed of a total of 16 residual blocks, with 4 blocks consisting of downsampling operation.

ResNet50: A 50-layer residual network composed of a total of 16 residual blocks, with 4 blocks consisting of downsampling operation.

ResNet101: A 101-layer residual network composing of a total of 33 residual blocks, with 4 blocks consisting of downsampling operation.

[1] https://huggingface.co/datasets/dpdl-benchmark/places365-mini-class-hard
[2] https://huggingface.co/datasets/timm/oxford-iiit-pet
[3] https://huggingface.co/datasets/zh-plus/tiny-imagenet
[4] https://huggingface.co/datasets/nkirschi/oxford-flowers
[5] https://huggingface.co/datasets/mteb/Caltech101

ResNet152: A 152-layer residual network composing of a total of 50 residual blocks, with 4 blocks consisting of downsampling operation (the first block in the $conv_2$ stage only has channel increment.

WideResNet101: A widened variant of ResNet with 101 layers, specifically *wide_resnet101_2*, doubles the channel width of the standard ResNet101 architecture. It consists of a total of 33 residual blocks, with 4 blocks consisting of downsampling operation.

ResNeXt101: A ResNeXt architecture with 101 layers, specifically *ResNeXt101-32×4d*, employs aggregated transformations with a cardinality of 32 and a bottleneck width of 4. It consists of a total of 33 residual blocks, with 4 blocks consisting of downsampling operation.

A.3 Training Setup and Hyperparameters

We carry out our analysis under a controlled training setup designed to evaluate robustness to label noise. Specifically, we introduce label noise by randomly flipping the labels of $2\%, 5\%, 10\%$ of the training set samples.

All models are trained using the Adam optimizer with a learning rate of 2×10^{-5}. Training is performed for 40–50 epochs, until 100% train accuracy is reached, when applicable. We use a batch size of 16 and optimize the standard cross-entropy loss throughout all experiments.

Furthermore, to demonstrate the robustness of our results—namely, that our findings are not specific to any single model–dataset combination—we evaluate our findings across multiple dataset–model pairs: (i) CIFAR10 with ResNet152; (ii) Places365 with ResNet50 and ResNet101; (iii) Oxford-IIIT Pet with ResNet34, and ResNet152; (iv) Oxford Flower 102 with ResNet50, ResNet101, and ResNeXt101; (v) TinyImageNet with ResNet101, and (vi) Caltech101 with ResNet152 and WideResNet101.

B Different Ablation Strategies Used

In this study, we utilize three different ablation strategies to understand how differently residual connections and main branch propagate memorization and generalization throughout the network, as follows:

Zero Ablation: (Veit et al., 2016; Bhojanapalli et al., 2021), replaces the model pathway's (residual connection/main branch) output with a zero vector, providing a simple baseline that removes all information carried by that pathway.

Mean Ablation: (Tigges et al., 2023; McDougall et al., 2023; Ghorbani & Zou, 2020) replaces the model pathway's (residual connection/main branch) output with a mean activation vector computed using a calibration set. We use the held-out validation set since the model has never been trained on it.

Optimal Ablation: (Li & Janson, 2024) replaces the model pathway's output with a learned constant $a^\star$, optimized to minimize the KL divergence between the ablated and original network output distributions over the calibration set. By finding the best possible constant substitution, it removes any spoofing impact due to any fixed activation replacement. For the calibration set, we use the held-out validation set, 5 epochs of finetuning (by freezing all model parameters), learning rate of 1×10^{-4}, and Adam optimizer, to obtain the learned constant $a^\star$ for each model pathway (residual connection and main branch) throughout the network.

C Additional Experiments and Results

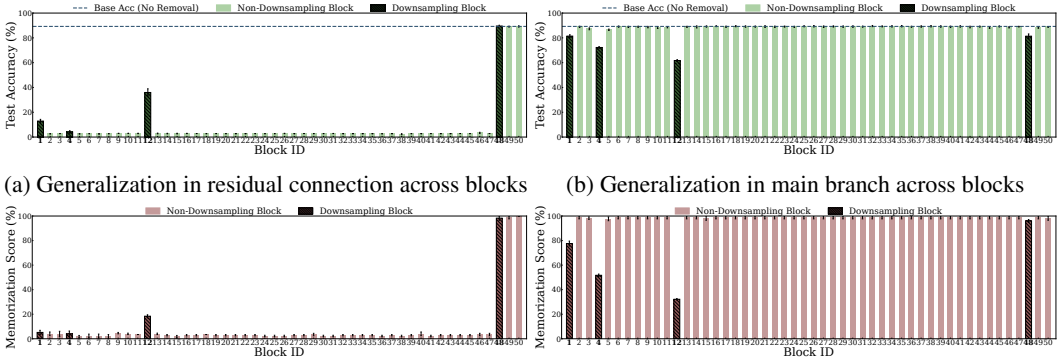
In this section, we present additional experimental results corresponding to the analyses discussed in Sections 4, 5, 6, 7, and 8, covering remaining interpretability strategies, models, and datasets. These results aim to: **(i)** show downsampling blocks and residual connections propagate memorization along with generalization, with main branch in non-downsampling blocks having minimal impact on both; **(ii)** verify the consistency of these observations under higher label noise levels; **(iii)**, verify the consistency of these observations under mean and optimal ablation interpretability approaches, **(iv)** understand the role of downsampling blocks in propagating memorization via fisher information trace; **(v)** understand how residual connections and main branch propagate memorization and generalization

using activation-level analysis; **(vi)** show memorization and generalization weights are primarily concentrated in downsampling blocks and are entangled together, and **(vii)** analyze the impact of ablating residual connections and the main branch on the model's test-time prediction confidence across blocks. The corresponding results are presented throughout Sections C.1, C.2, C.3, C.4, C.5, C.6, and C.7.

C.1 Impact of Downsampling and Non-Downsampling Residual Blocks on Memorization and Generalization

In this section, we provide additional results examining how ablating residual connections and the main branch affect memorization and generalization across downsampling and non-downsampling blocks in the ResNet architecture. These findings further corroborate our central observation that downsampling blocks and residual connections propagate memorization along with generalization.

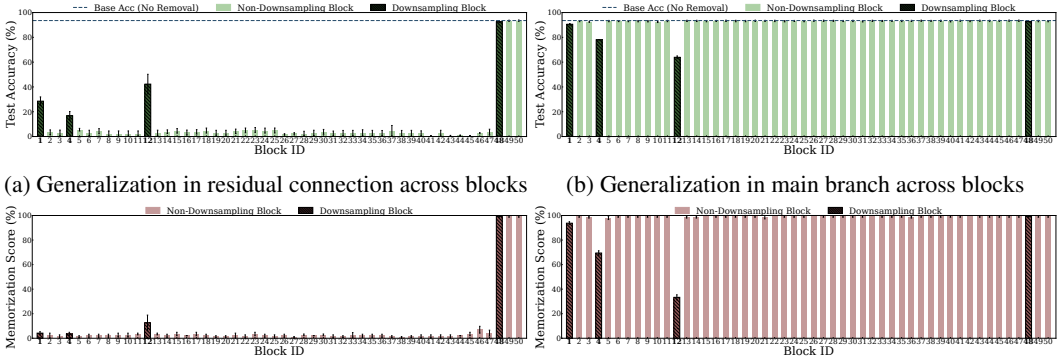
Specifically, we show that both residual connections and the main branch propagate memorization and generalization within downsampling blocks, whereas residual connections in non-downsampling blocks are the primary carriers for both information. These results are validated across multiple datasets—CIFAR10, Places365, Oxford Flower-102, Oxford-IIIT Pet, TinyImagenet, and Caltech101—and model architectures—ResNet34, ResNet50, ResNet101, ResNet152, WideResNet101, and ResNeXt101—as shown in Figs. 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18.



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

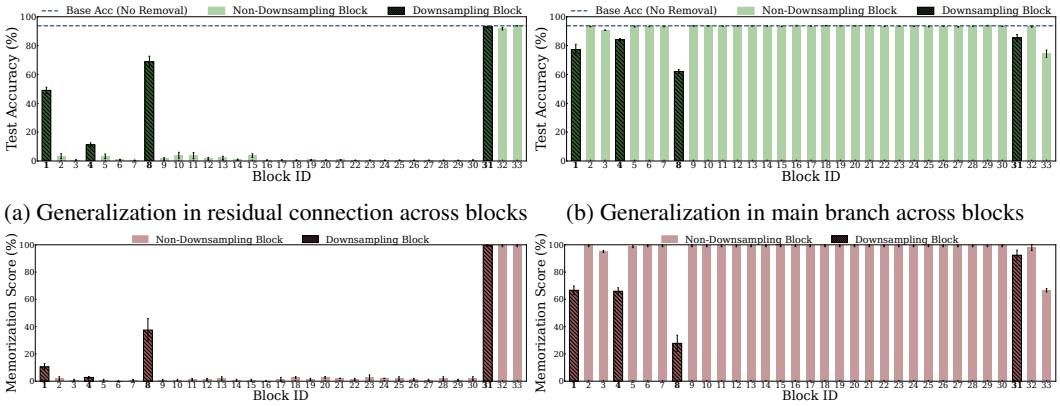
Figure 9: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (Oxford IIIT-Pet, ResNet-152, Zero Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

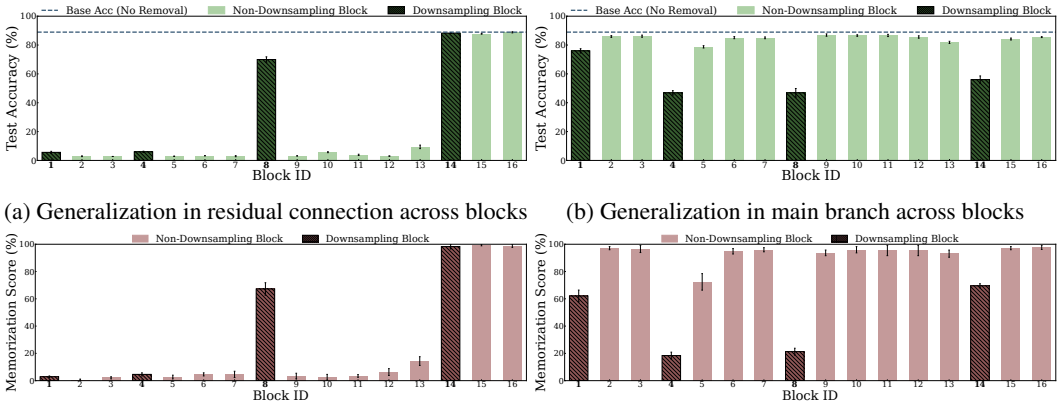
Figure 10: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (Caltech101, ResNet-152, Zero Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

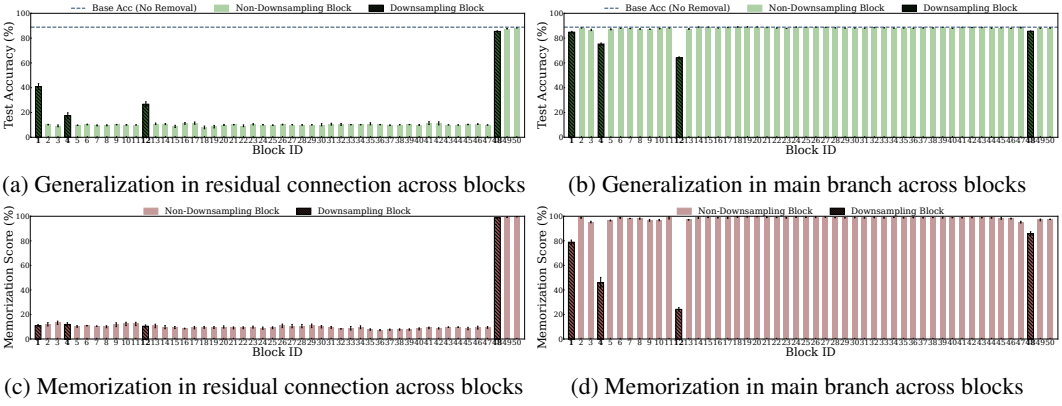
Figure 11: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Caltech101, WideResNet101, Zero Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

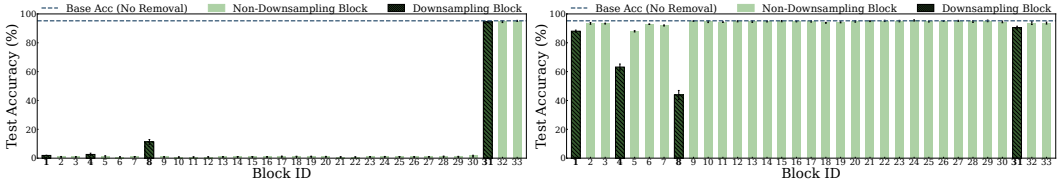
Figure 12: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 15, 16) have a negligible impact on both memorization and generalization (Oxford IIIT-Pet, ResNet34, Zero Ablation). We note that in smaller ResNets (like ResNet34/50), memorization and generalization become slightly dispersed across the downsampling and non-downsampling blocks main branch. However, the downsampling blocks still exhibit a distinct concentration of memorization. We attribute this dispersion primarily to the lower parameter count (smaller model size), which causes a smaller model to utilize more blocks than a larger model to memorize noisy labels.



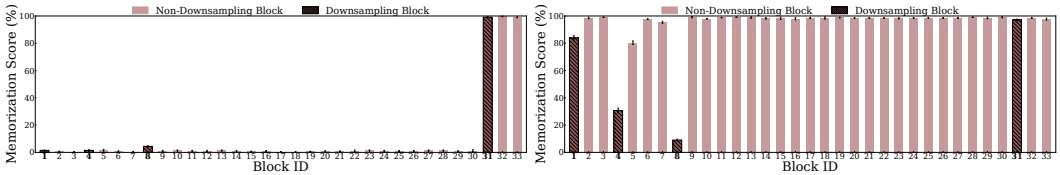
(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 13: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (CIFAR-10, ResNet-152, Zero Ablation).

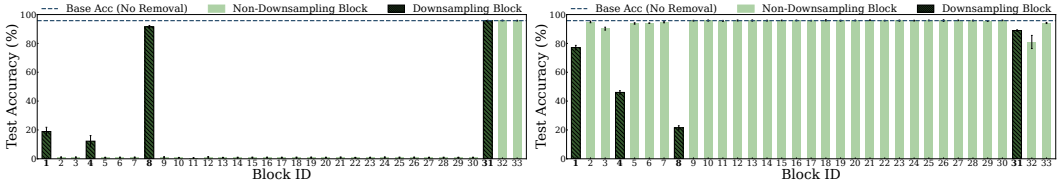


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

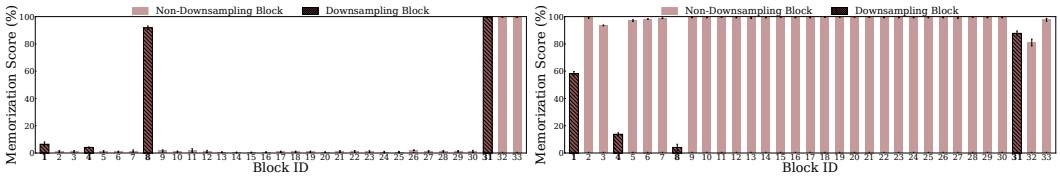


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 14: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Oxford Flower-102, ResNet-101, Zero Ablation).

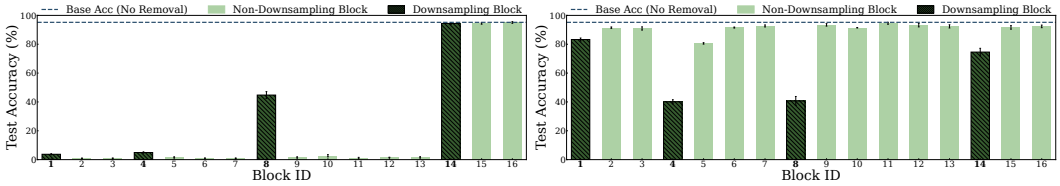


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

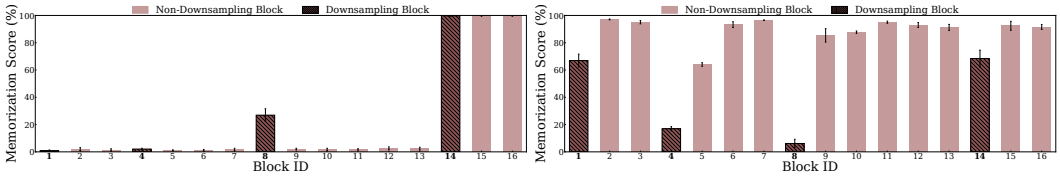


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 15: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Oxford Flower-102, ResNeXt-101, Zero Ablation).

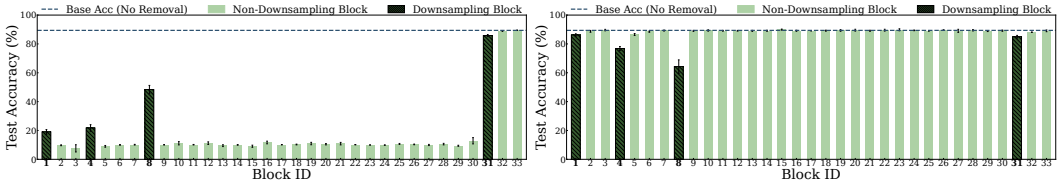


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

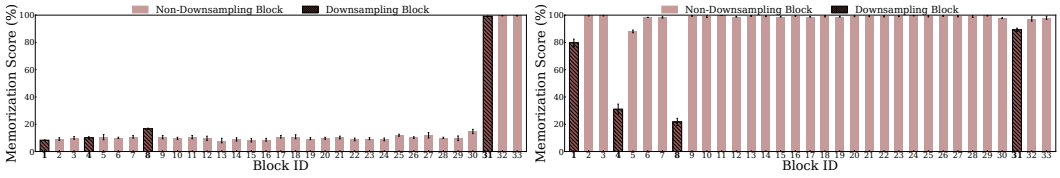


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 16: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (Oxford Flower-102, ResNet-50, Zero Ablation). We note that in smaller ResNets (like ResNet34/50), memorization and generalization become slightly dispersed across the downsampling and non-downsampling blocks main branch. However, the downsampling blocks still exhibit a distinct concentration of memorization. We attribute this dispersion primarily to the lower parameter count (smaller model size), which causes a smaller model to utilize more blocks than a larger model to memorize noisy labels.

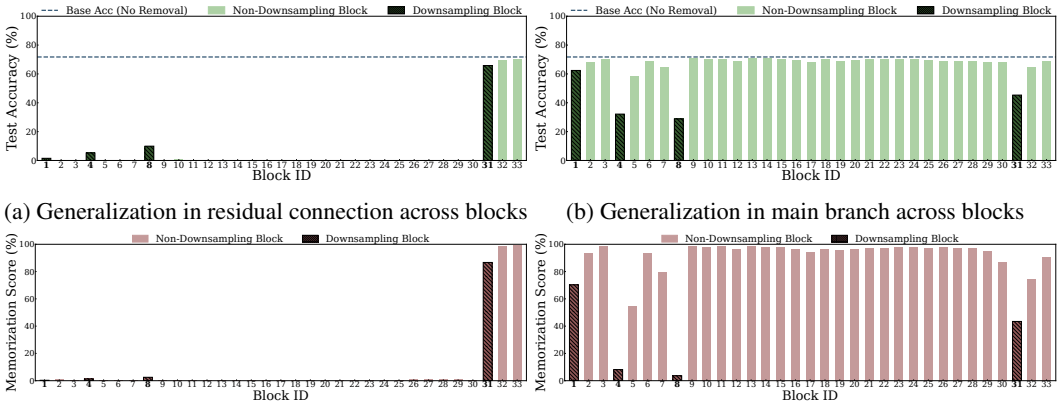


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks



(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 17: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Places365, ResNet-101, Zero Ablation).



(a) Generalization in residual connection across blocks

(b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks

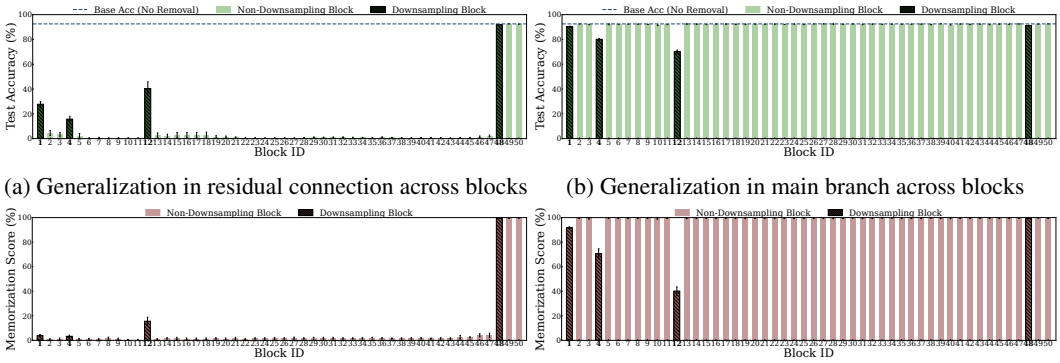
(d) Memorization in main branch across blocks

Figure 18: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the residual connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (TinyImagenet, ResNet-101, Zero Ablation).

C.2 Consistency Across Higher Noise Levels

To further assess the robustness of our findings, we evaluate the impact of residual connections and main branch in downsampling and non-downsampling blocks on memorization and generalization, under higher label noise levels of 5% and 10%, extending the 2% noise setting presented in Appendix C.1. Across all datasets—CIFAR10, Places365, Oxford Flower-102, Oxford-IIIT Pet, and Caltech101—and model architectures—ResNet34, ResNet50, ResNet101, ResNet152, WideResNet101, and ResNeXt101—we observe consistent trends that reinforce our earlier conclusions. The corresponding results are presented in Figs. 19, 20, 21, 22, 23, 24, 25, 3 for 5% label noise and Figs. 26, 27, 28, 29, 30, 31, 32, 33 for 10% label noise.

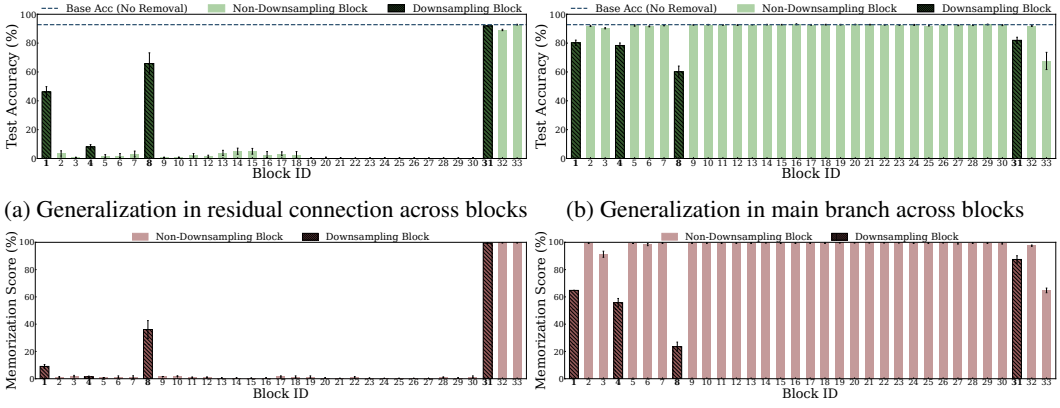
C.2.1 5% Label Noise Results



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

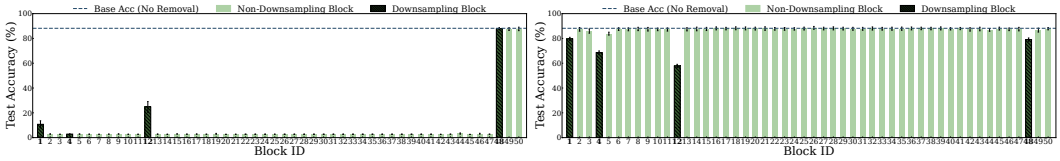
Figure 19: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Converself, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Caltech101, ResNet-152, 5% Label Noise, Zero Ablation).



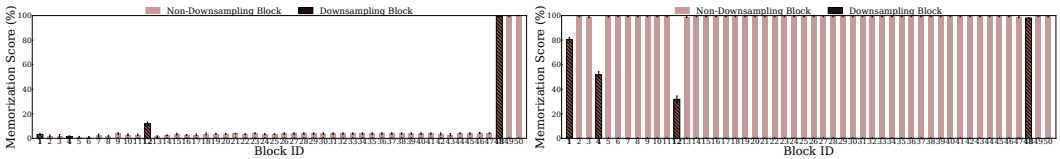
(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 20: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Converself, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Caltech101, WideResNet-101, 5% Label Noise, Zero Ablation).

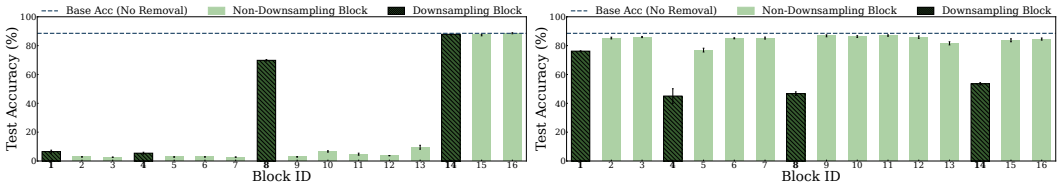


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

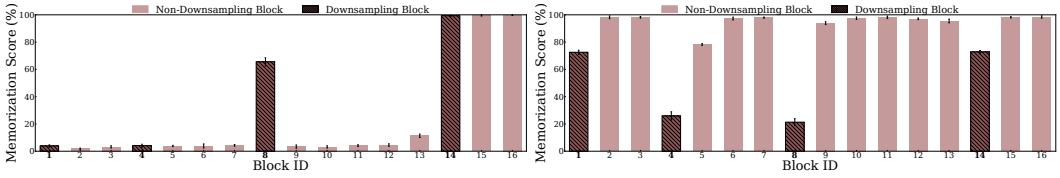


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 21: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Converself, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford IIIT-Pet, ResNet-152, 5% Label Noise, Zero Ablation).

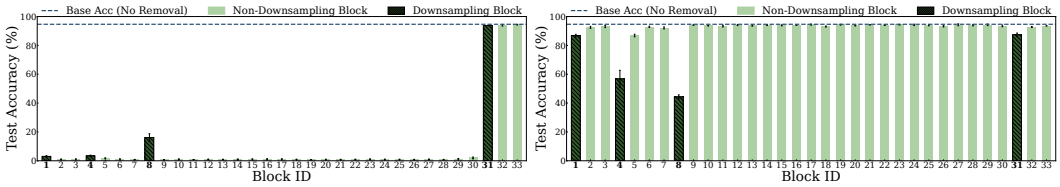


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

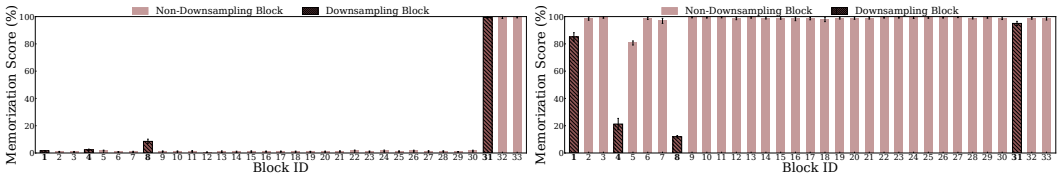


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 22: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Converself, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford IIIT-Pet, ResNet-34, 5% Label Noise, Zero Ablation). We note that in smaller ResNets (like ResNet34/50), memorization and generalization become slightly dispersed across the downsampling and non-downsampling blocks main branch. However, the downsampling blocks still exhibit a distinct concentration of memorization. We attribute this dispersion primarily to the lower parameter count (smaller model size), which causes a smaller model to utilize more blocks than a larger model to memorize noisy labels.

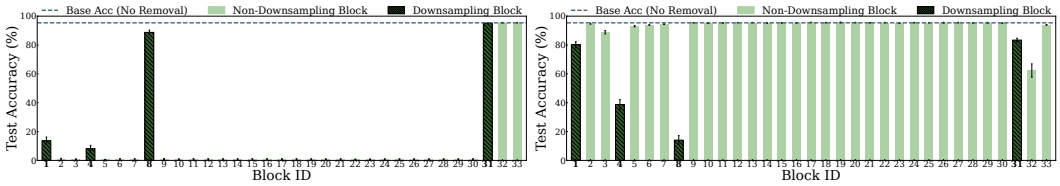


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

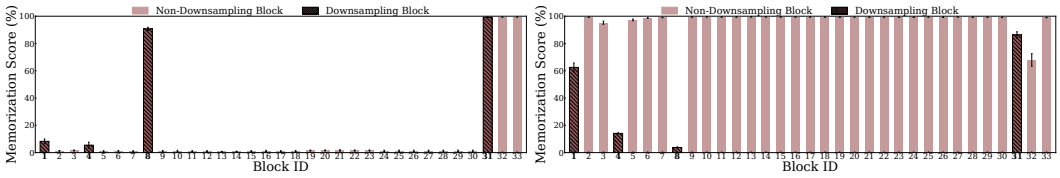


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 23: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Converself, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford Flower-102, ResNet-101, 5% Label Noise, Zero Ablation).

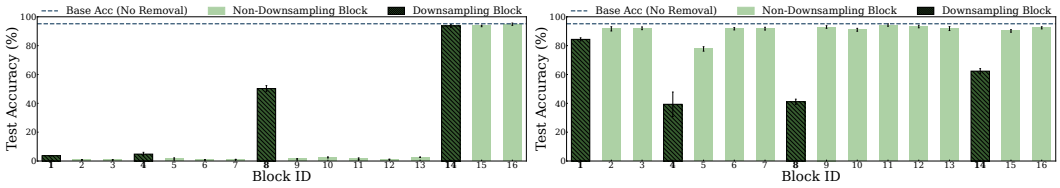


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks



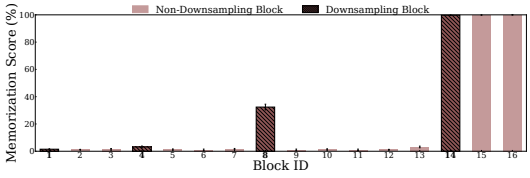
(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 24: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Converself, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford Flower-102, ResNeXt-101, 5% Label Noise, Zero Ablation).

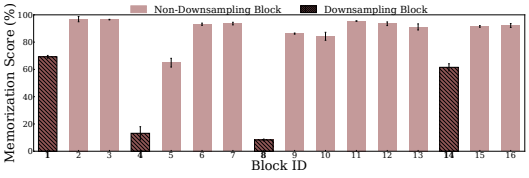


(a) Generalization in residual connection across blocks

(b) Generalization in main branch across blocks



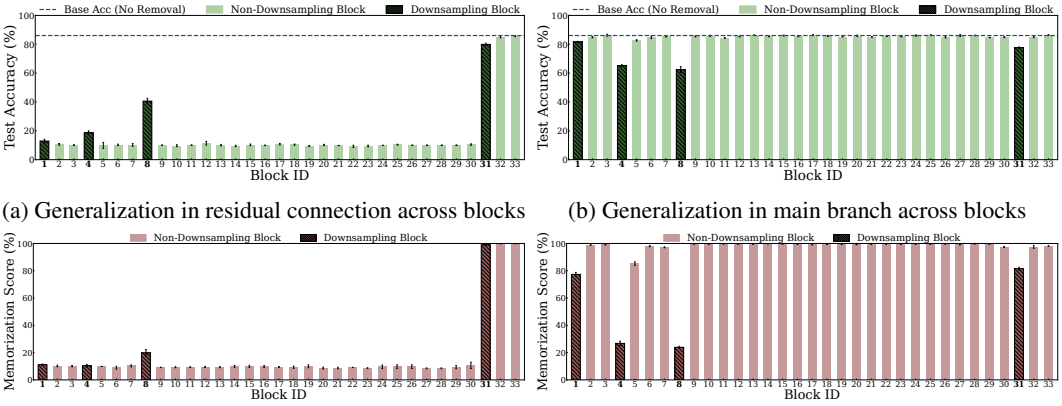
(c) Memorization in residual connection across blocks



(d) Memorization in main branch across blocks

Figure 25: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Conversely, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford Flower-102, ResNet-50, 5% Label Noise, Zero Ablation). We note that in smaller ResNets (like ResNet34/50), memorization and generalization become slightly dispersed across the downsampling and non-downsampling blocks in main branch. However, the downsampling blocks still exhibit a distinct concentration of memorization. We attribute this dispersion primarily to the lower parameter count (smaller model size), which causes a smaller model to utilize more blocks than a larger model to memorize noisy labels.

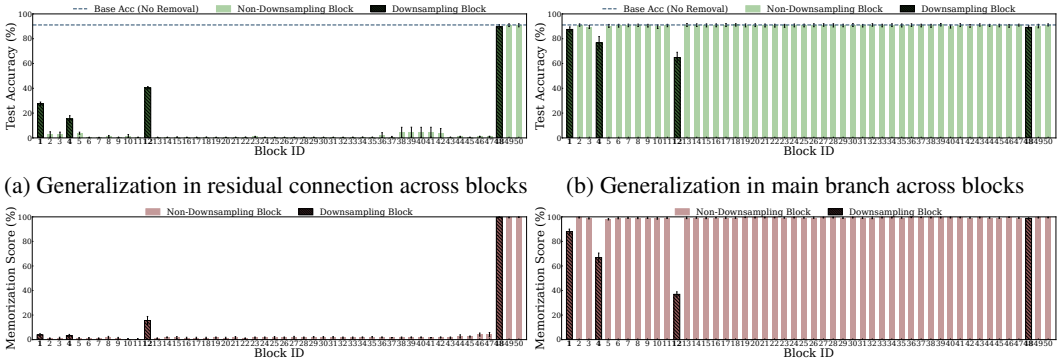
C.2.2 10% LABEL NOISE RESULTS



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

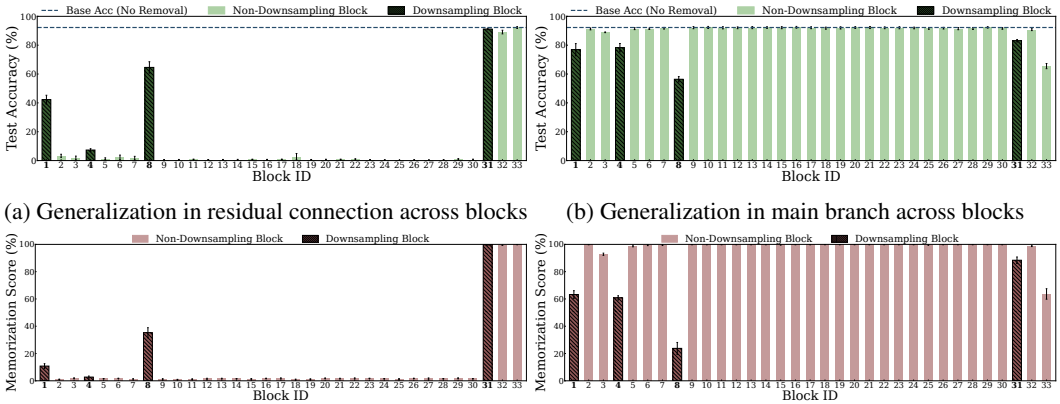
Figure 26: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Conversely, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Places365, ResNet-101, 10% Label Noise, Zero Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

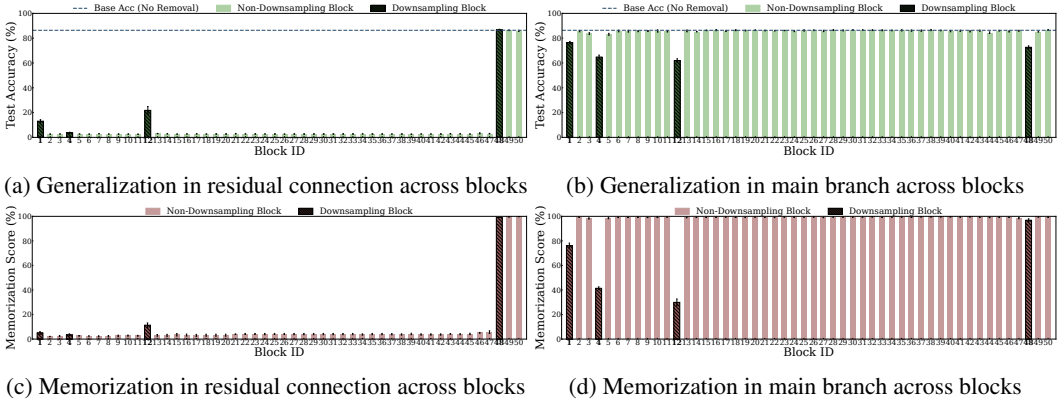
Figure 27: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Conversely, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Caltech101, ResNet-152, 10% Label Noise, Zero Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

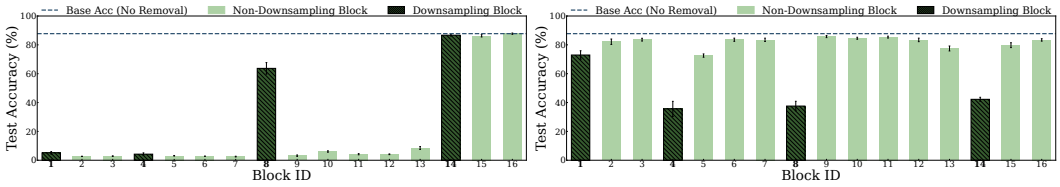
Figure 28: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Conversely, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Caltech101, WideResNet-101, 10% Label Noise, Zero Ablation).



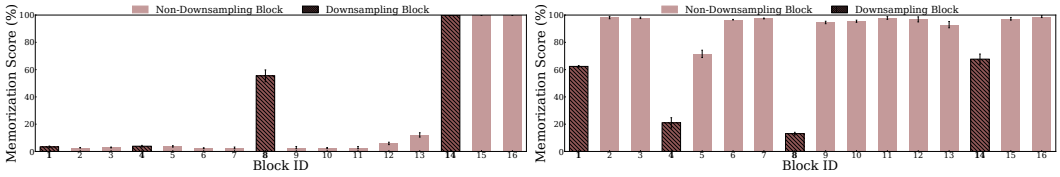
(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 29: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Conversely, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford IIIT-Pet, ResNet-152, 10% Label Noise, Zero Ablation).

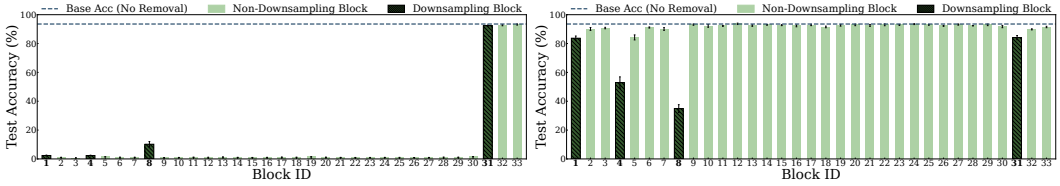


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

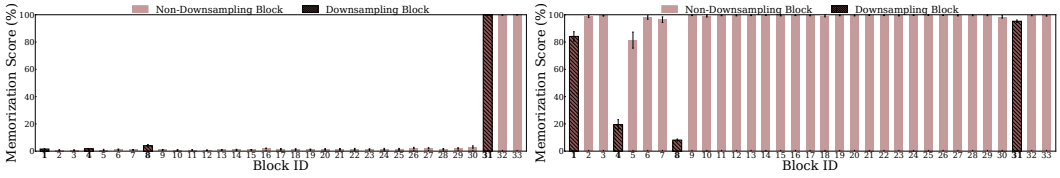


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 30: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Conversely, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford IIIT-Pet, ResNet-34, 10% Label Noise, Zero Ablation). We note that in smaller ResNets (like ResNet34/50), memorization and generalization become slightly dispersed across the downsampling and non-downsampling blocks main branch. However, the downsampling blocks still exhibit a distinct concentration of memorization. We attribute this dispersion primarily to the lower parameter count (smaller model size), which causes a smaller model to utilize more blocks than a larger model to memorize noisy labels.

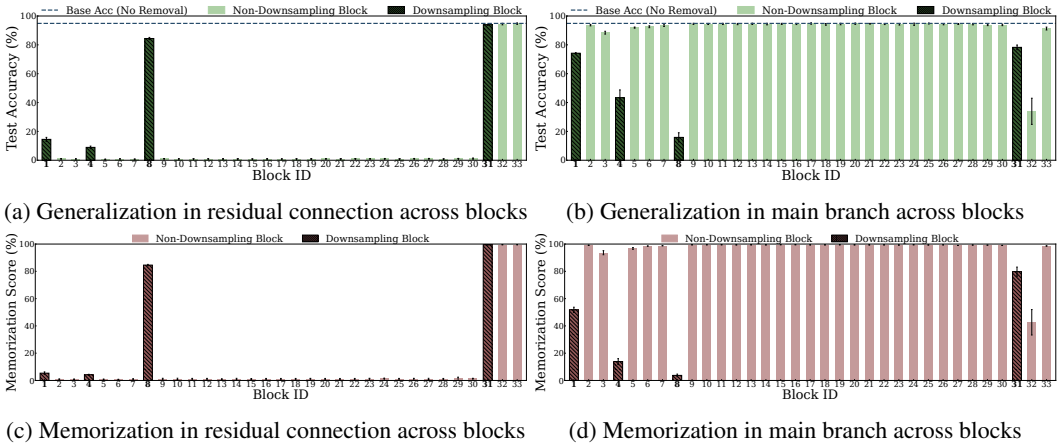


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks



(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

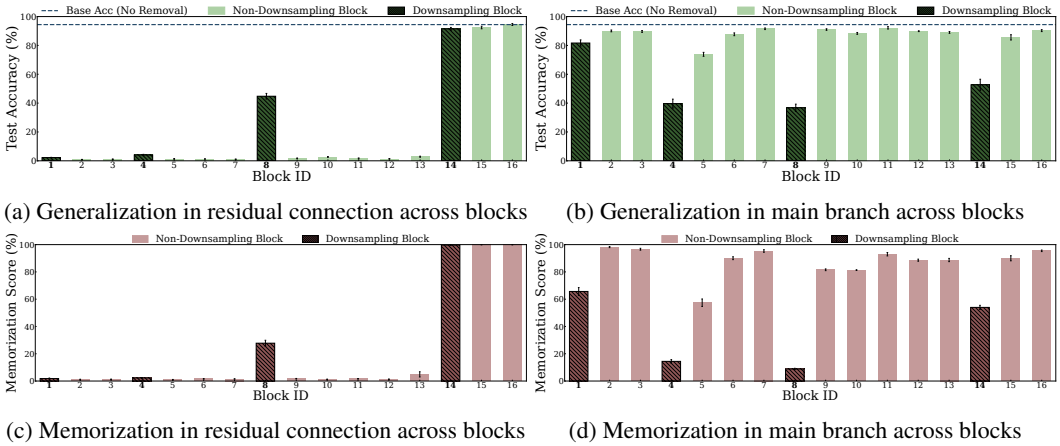
Figure 31: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Conversely, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford Flower-102, ResNet-101, 10% Label Noise, Zero Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 32: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Conversely, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford Flower-102, ResNeXt-101, 10% Label Noise, Zero Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

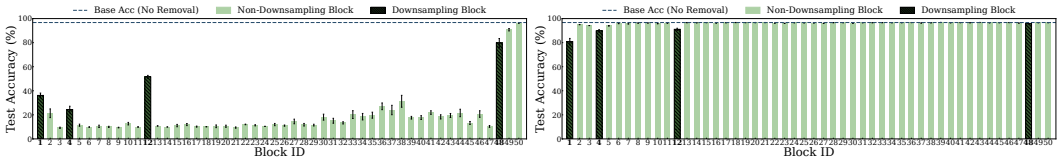
(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 33: **Robust trends persist under higher label noise.** Repeating the analysis at higher label noise levels reveals the same qualitative behavior: Residual connections and main branch in downsampling blocks propagate memorization and generalization. Conversely, residual connections in non-downsampling blocks are the primary carriers for both, with the main branch having minimal impact (Oxford Flower-102, ResNet-50, 10% Label Noise, Zero Ablation). We note that in smaller ResNets (like ResNet34/50), memorization and generalization become slightly dispersed across the downsampling and non-downsampling blocks main branch. However, the downsampling blocks still exhibit a distinct concentration of memorization. We attribute this dispersion primarily to the lower parameter count (smaller model size), which causes a smaller model to utilize more blocks than a larger model to memorize noisy labels.

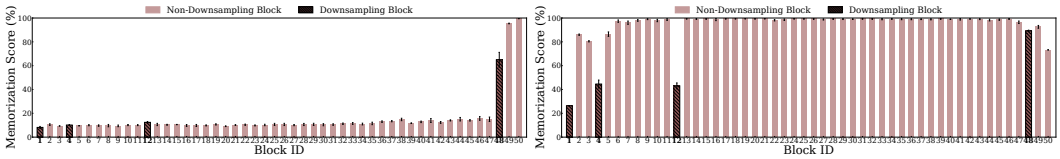
C.3 Consistent Results across Optimal and Mean Ablation Interpretability Approaches

In this section, we provide results corresponding to optimal and mean ablation to show that our claims on downsampling blocks and residual connections propagating memorization, along with generalization, remain consistent regardless of the ablation strategy used. The results are presented across multiple datasets—CIFAR10, Places365, Oxford Flower-102, Oxford-IIIT Pet, and Caltech101—and architectures—ResNet34, ResNet50, ResNet101, ResNet152, WideResNet101, and ResNeXt101, in Sections C.3.1 & C.3.2.

C.3.1 Optimal Ablation Results

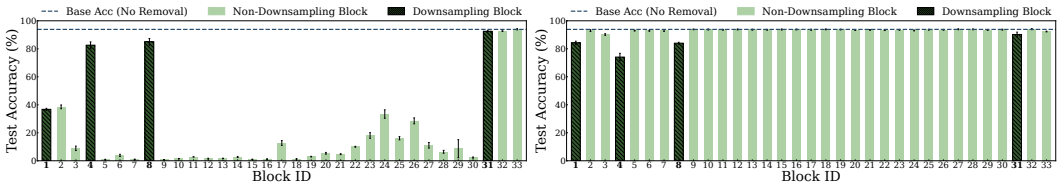


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

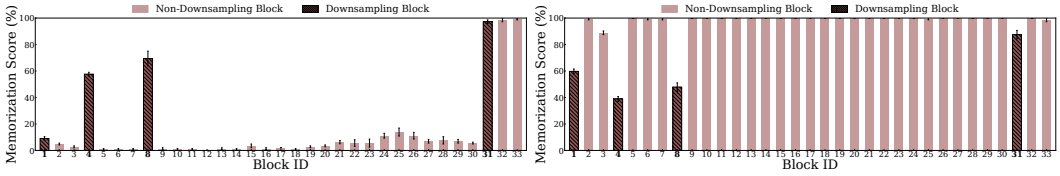


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 34: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (CIFAR10, ResNet-152, Optimal Ablation).

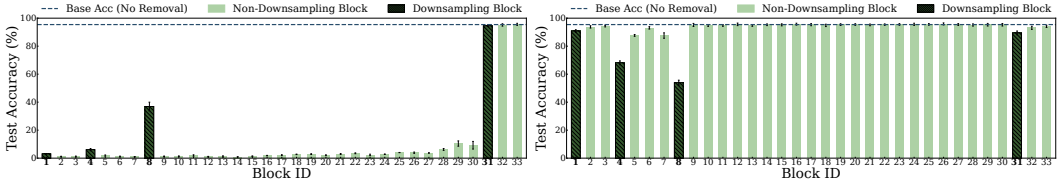


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

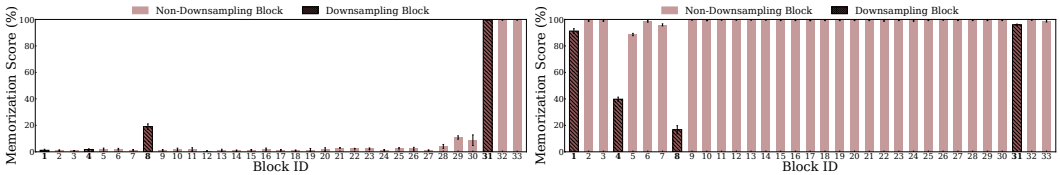


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 35: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Caltech101, WideResNet-101, Optimal Ablation).

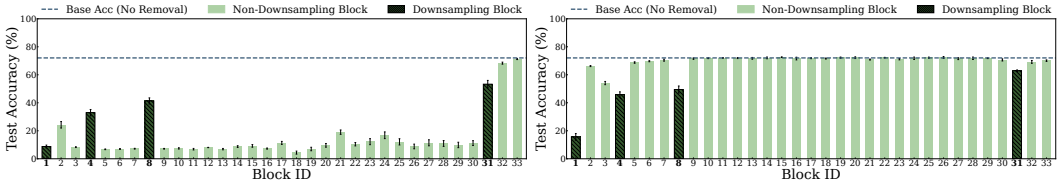


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

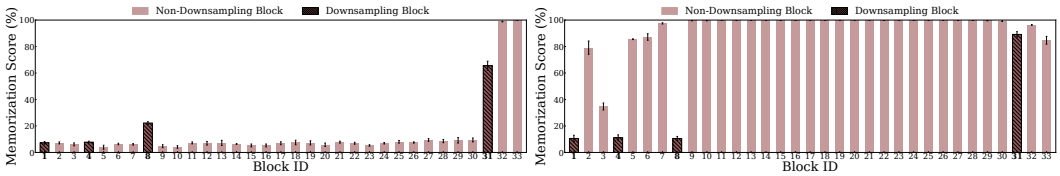


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 36: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Oxford Flower-102, ResNet-101, Optimal Ablation).

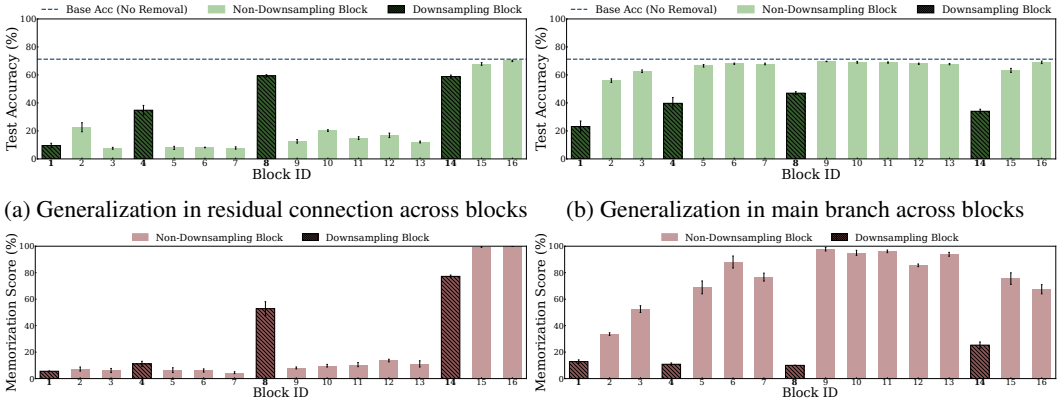


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks



(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

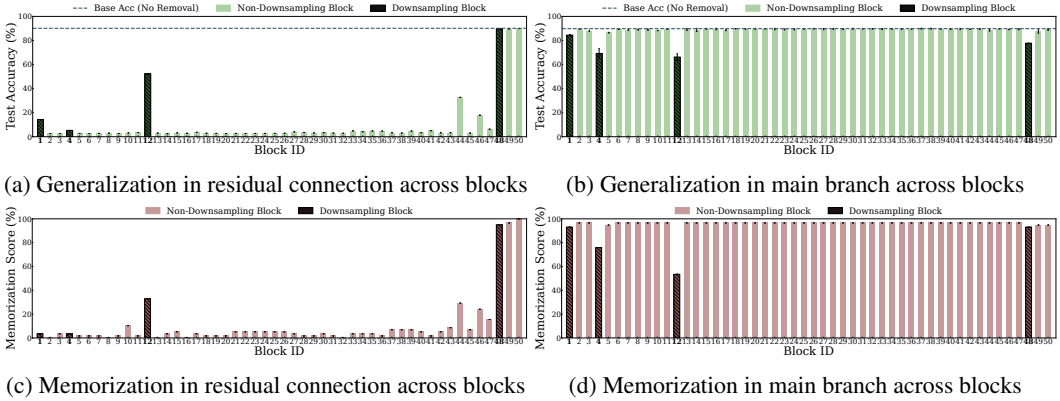
Figure 37: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Places365, ResNet-101, Optimal Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

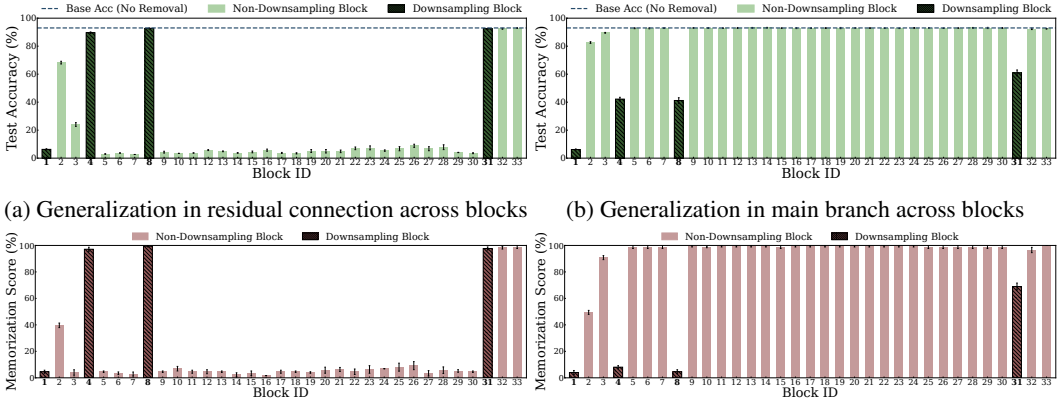
Figure 38: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 15, 16) have a negligible impact on both memorization and generalization (Places365, ResNet-50, Optimal Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

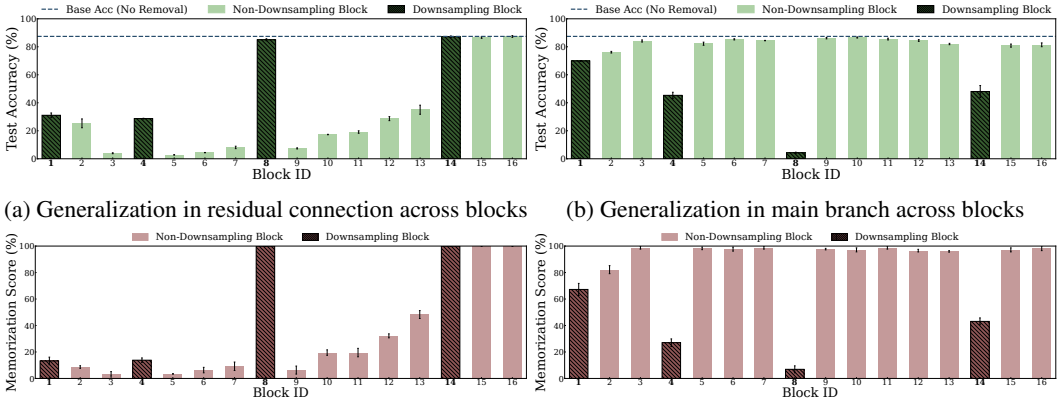
Figure 39: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (Oxford IIIT-Pet, ResNet-152, Optimal Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 40: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Oxford IIIT-Pet, ResNeXt-101, Optimal Ablation).

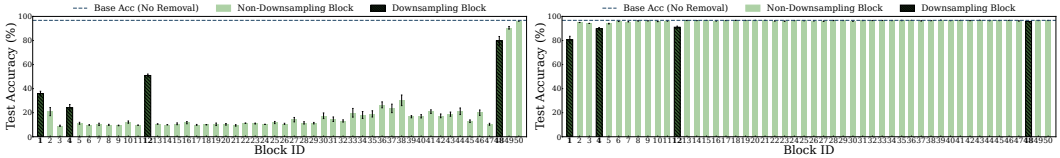


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

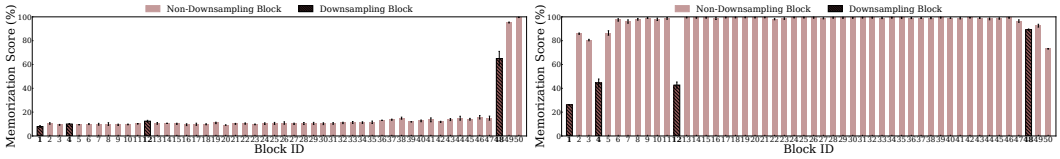
(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 41: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 15, 16) have a negligible impact on both memorization and generalization (Oxford IIIT-Pet, ResNet-34, Optimal Ablation).

C.3.2 MEAN ABLATION RESULTS

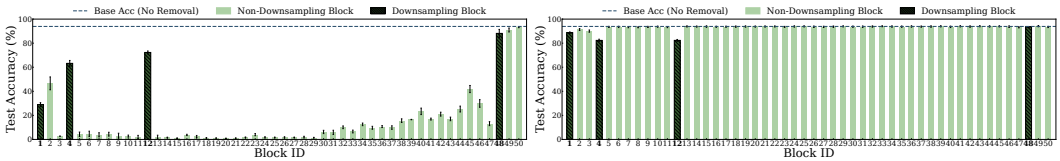


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

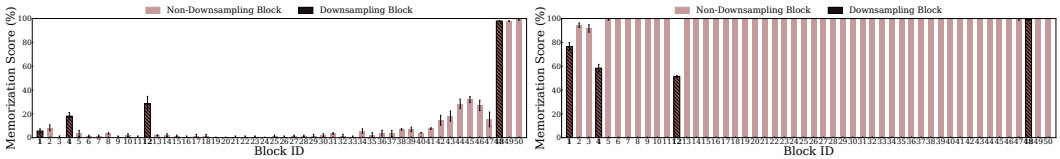


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 42: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (CIFAR10, ResNet-152, Mean Ablation).

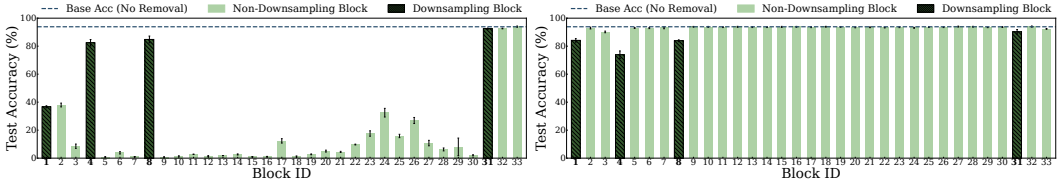


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

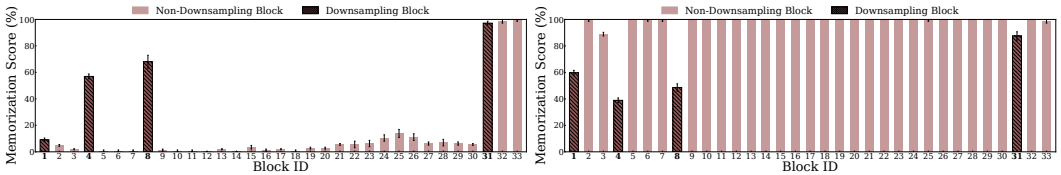


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 43: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (Caltech101, ResNet-152, Mean Ablation).

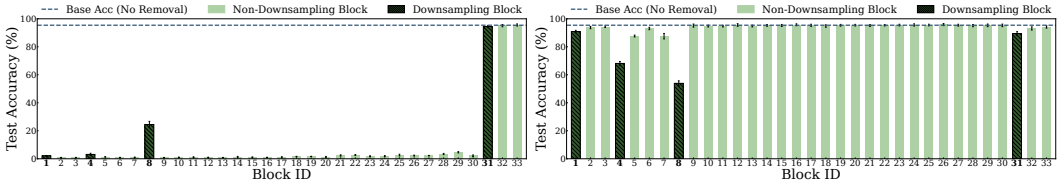


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

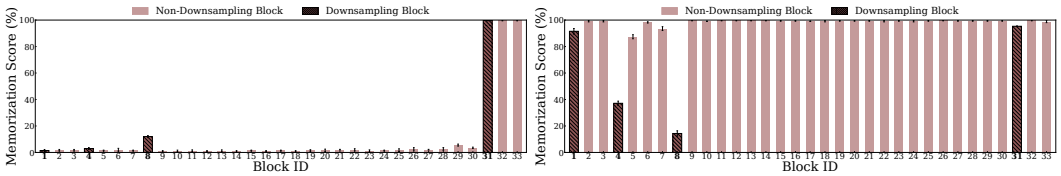


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 44: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Caltech101, WideResNet-101, Mean Ablation).

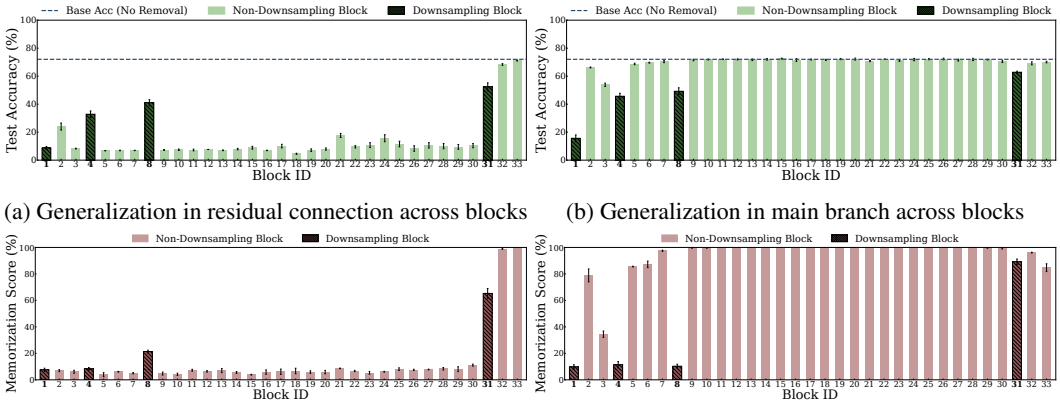


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks



(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

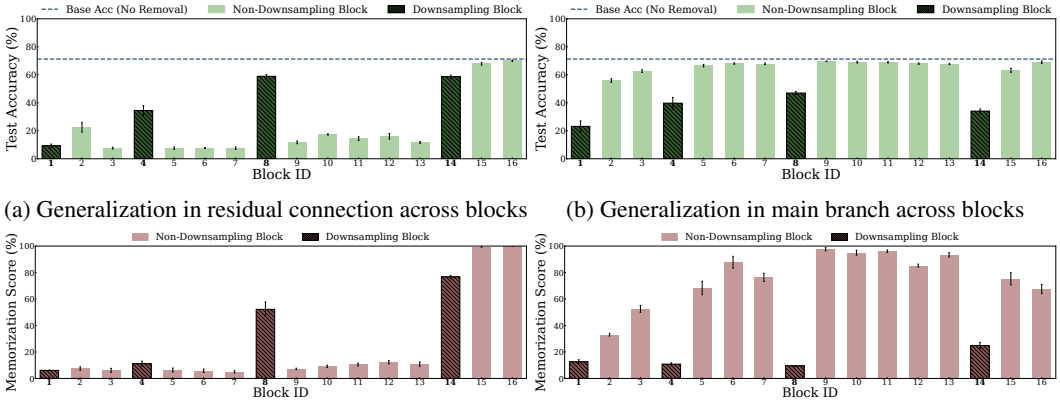
Figure 45: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Oxford Flower-102, ResNet-101, Mean Ablation).



(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

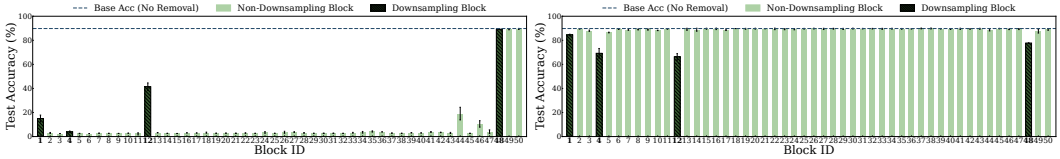
Figure 46: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Places365, ResNet-101, Mean Ablation).



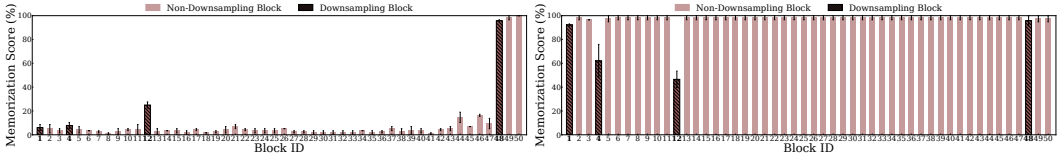
(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 47: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 15, 16) have a negligible impact on both memorization and generalization (Places365, ResNet-50, Mean Ablation).

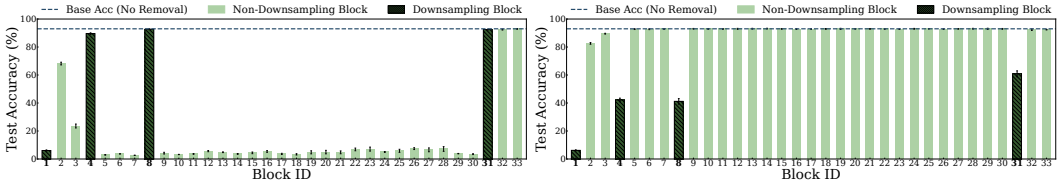


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks

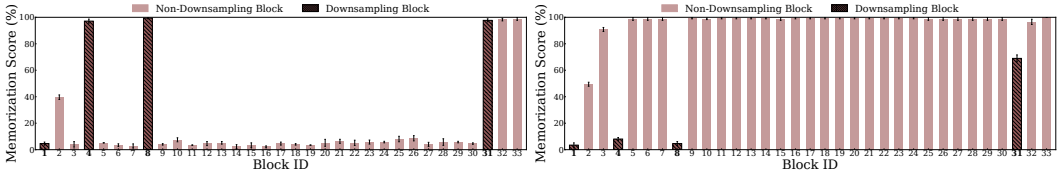


(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 48: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 49, 50) have a negligible impact on both memorization and generalization (Oxford IIIT-Pet, ResNet-152, Mean Ablation).

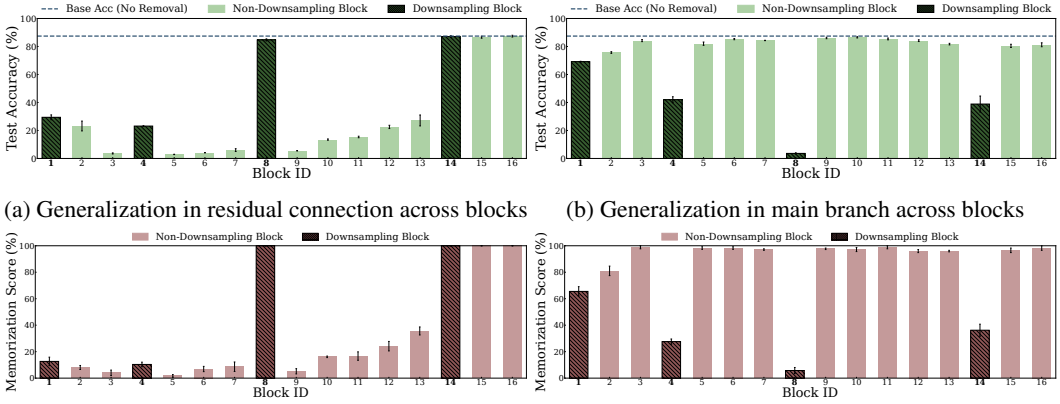


(a) Generalization in residual connection across blocks (b) Generalization in main branch across blocks



(c) Memorization in residual connection across blocks (d) Memorization in main branch across blocks

Figure 49: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 32, 33) have a negligible impact on both memorization and generalization (Oxford IIIT-Pet, ResNeXt-101, Mean Ablation).



(a) Generalization in residual connection across blocks

(b) Generalization in main branch across blocks

(c) Memorization in residual connection across blocks

(d) Memorization in main branch across blocks

Figure 50: **Downsampling blocks and residual connections propagate memorization along with generalization.** Residual connections and main branch in downsampling blocks propagate memorization and generalization. In contrast, residual connections in non-downsampling blocks are the primary carriers for both phenomena. One more point to note is that the skip connections in the last two non-downsampling blocks (Block IDs 15, 16) have a negligible impact on both memorization and generalization (Oxford IIIT-Pet, ResNet-34, Mean Ablation).

C.4 Fisher Information Trace Tr(F) Analysis

In this section, we provide additional results illustrating the relationship between the Fisher information trace Tr(F) and memorization across residual blocks. We show that downsampling blocks consistently exhibit higher Tr(F) compared to non-downsampling blocks, providing an intuitive understanding of the dominant impact of downsampling blocks on memorization. Results across multiple datasets—CIFAR10, Places365, Oxford Flower-102, Oxford-IIIT Pet, and Caltech101—and architectures—ResNet34, ResNet50, ResNet101, ResNet152, WideResNet101, and ResNeXt101—are presented in Figs. 5, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70.

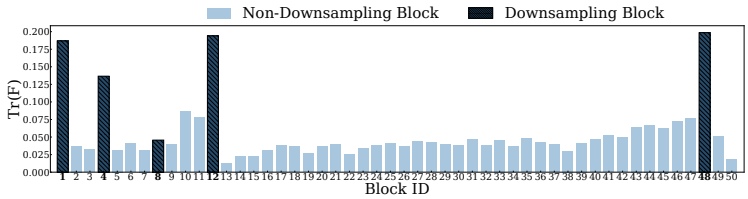


Figure 51: Trace of Fisher Information Matrix, Tr(F), across blocks (CIFAR10-ResNet152)

Figure 52: **Downsampling blocks exhibit higher Tr(F) than non-downsampling blocks.** We compute the block-wise Fisher information trace Tr(F) and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization.

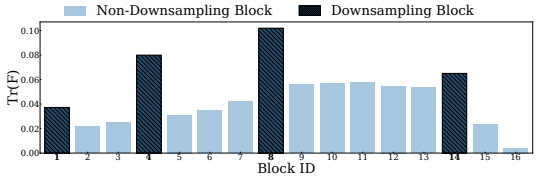


Figure 53: Trace of Fisher Information Matrix, Tr(F), across blocks (CIFAR10-ResNet34)

Figure 54: **Downsampling blocks exhibit higher Tr(F) than non-downsampling blocks.** We compute the block-wise Fisher information trace Tr(F) and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization. Note that for smaller models (like ResNet34/50), Tr(F) seems to be comparable across downsampling and non-downsampling blocks, further confirming previous observation that for smaller ResNets memorization seems to get slightly more dispersed across downsampling and non-downsampling blocks.

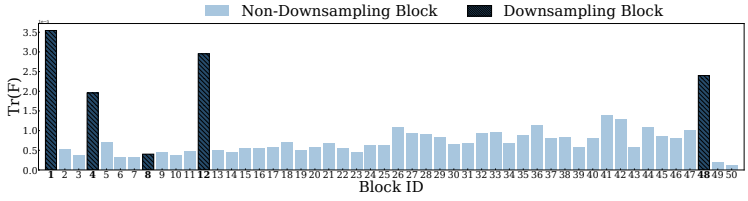


Figure 55: Trace of Fisher Information Matrix, Tr(F), across blocks (Caltech101-ResNet152)

Figure 56: **Downsampling blocks exhibit higher Tr(F) than non-downsampling blocks.** We compute the block-wise Fisher information trace Tr(F) and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization.

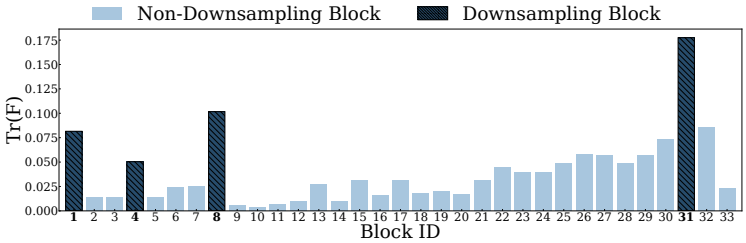


Figure 57: Trace of Fisher Information Matrix, $\mathrm{Tr}(\mathrm{F})$, across blocks (Caltech101-WideResNet101)

Figure 58: **Downsampling blocks exhibit higher $\mathrm{Tr}(\mathbf{F})$ than non-downsampling blocks.** We compute the block-wise Fisher information trace $\mathrm{Tr}(\mathrm{F})$ and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization.

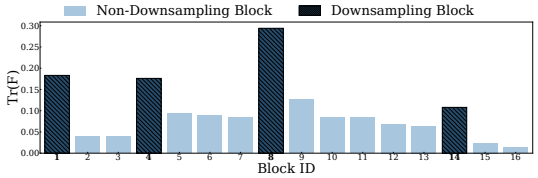


Figure 59: Trace of Fisher Information Matrix, $\mathrm{Tr}(\mathrm{F})$, across blocks (Caltech101-ResNet50)

Figure 60: **Downsampling blocks exhibit higher $\mathrm{Tr}(\mathbf{F})$ than non-downsampling blocks.** We compute the block-wise Fisher information trace $\mathrm{Tr}(\mathrm{F})$ and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization.

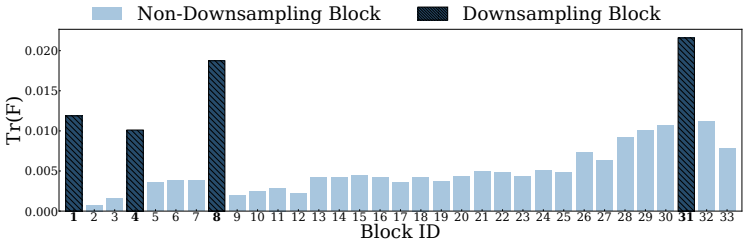


Figure 61: Trace of Fisher Information Matrix, $\mathrm{Tr}(\mathrm{F})$, across blocks (Oxford Flower-102-ResNet101)

Figure 62: **Downsampling blocks exhibit higher $\mathrm{Tr}(\mathbf{F})$ than non-downsampling blocks.** We compute the block-wise Fisher information trace $\mathrm{Tr}(\mathrm{F})$ and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization.

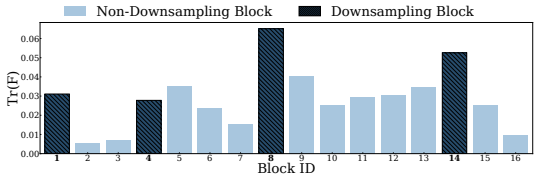


Figure 63: Trace of Fisher Information Matrix, $\mathrm{Tr}(\mathrm{F})$, across blocks (Oxford Flower-102-ResNet50)

Figure 64: **Downsampling blocks exhibit higher $\mathrm{Tr}(\mathbf{F})$ than non-downsampling blocks.** We compute the block-wise Fisher information trace $\mathrm{Tr}(\mathrm{F})$ and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization. Note that for smaller models (like ResNet34/50), $\mathrm{Tr}(\mathrm{F})$ seems to be comparable across downsampling and non-downsampling blocks, further confirming previous observation that for smaller ResNets memorization seems to get slightly more dispersed across downsampling and non-downsampling blocks.

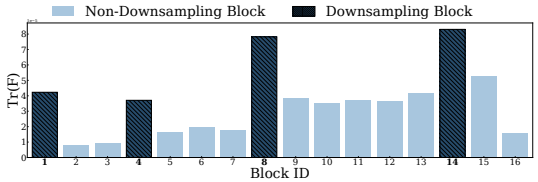


Figure 65: Trace of Fisher Information Matrix, $\mathrm{Tr}(\mathrm{F})$, across blocks (Places365-ResNet50)

Figure 66: **Downsampling blocks exhibit higher $\mathrm{Tr}(\mathbf{F})$ than non-downsampling blocks.** We compute the block-wise Fisher information trace $\mathrm{Tr}(\mathrm{F})$ and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization. Note that for smaller models (like ResNet34/50), $\mathrm{Tr}(\mathrm{F})$ seems to be comparable across downsampling and non-downsampling blocks, further confirming previous observation that for smaller ResNets memorization seems to get slightly more dispersed across downsampling and non-downsampling blocks.

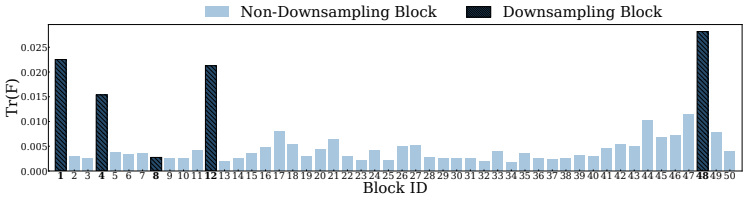


Figure 67: Trace of Fisher Information Matrix, $\mathrm{Tr}(\mathrm{F})$, across blocks (Oxford IIIT-Pet-ResNet152)

Figure 68: **Downsampling blocks exhibit higher $\mathrm{Tr}(\mathbf{F})$ than non-downsampling blocks.** We compute the block-wise Fisher information trace $\mathrm{Tr}(\mathrm{F})$ and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization.

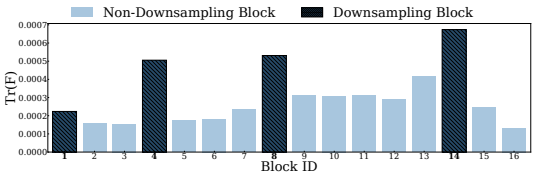
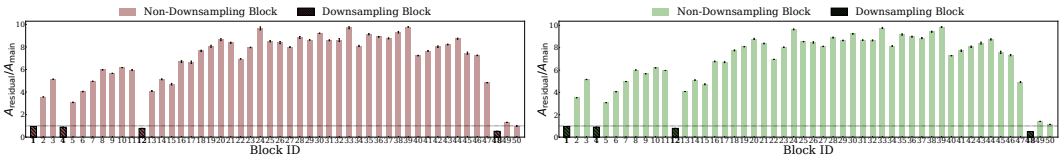


Figure 69: Trace of Fisher Information Matrix, $\mathrm{Tr}(\mathrm{F})$, across blocks (Oxford IIIT-Pet-ResNet34)

Figure 70: **Downsampling blocks exhibit higher $\mathrm{Tr}(\mathbf{F})$ than non-downsampling blocks.** We compute the block-wise Fisher information trace $\mathrm{Tr}(\mathrm{F})$ and see that downsampling blocks consistently exhibit much higher values than non-downsampling blocks, suggesting sharper loss curvature and hence a higher tendency for memorization. Note that for smaller models (like ResNet34/50), $\mathrm{Tr}(\mathrm{F})$ seems to be comparable across downsampling and non-downsampling blocks, further confirming previous observation that for smaller ResNets memorization seems to get slightly more dispersed across downsampling and non-downsampling blocks.

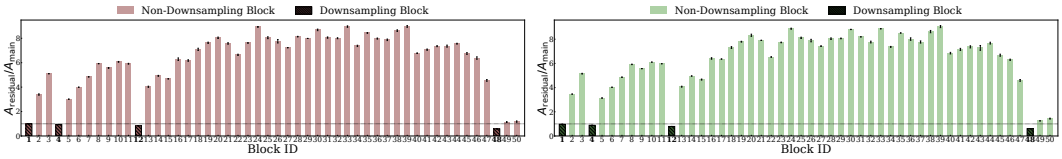
C.5 Activation-Level Analysis across Residual Blocks and Model Pathways

In this section, we provide further understanding on how differently residual connections and main branch propagate memorization and generalization across downsampling and non-downsampling blocks based on the activation-level analysis experiment discussed in Section 5. We corroborate our results using multiple datasets-CIFAR10, Places365, Oxford Flower-102, Oxford-IIIT Pet, and Caltech101, and architectures—ResNet34, ResNet50, ResNet101, ResNet152, WideResNet101, and ResNeXt101, presented in Figs. 4, 71, 72, 73, 74, 75, 76, and 77.



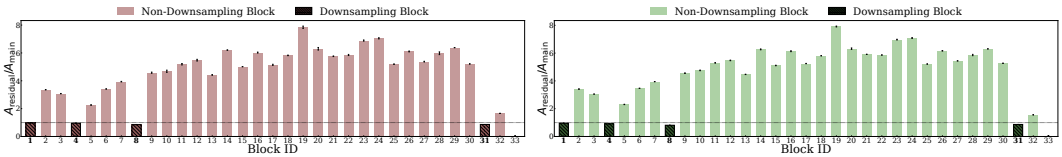
(a) Activation Analysis for Memorization (b) Activation Analysis for Generalization

Figure 71: **Activation-level analysis to understand how different model pathways propagate memorization and generalization.** We observe that at downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ depicting that both residual connections and main branch are propagating memorization and generalization. Conversely, in most non-downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \gg 1$, showing that residual connections are the primary carriers. Furthermore, in the last two non-downsampling blocks (Block IDs 49, 50) $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ confirming the observation made in Section 4.2 that residual connections and main branch in those blocks have a similar minimal impact on memorization and generalization (Oxford-IIIT Pet-ResNet152).



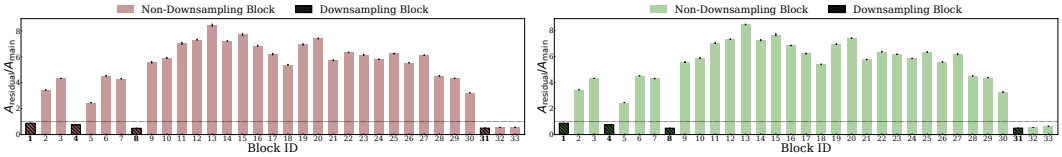
(a) Activation Analysis for Memorization (b) Activation Analysis for Generalization

Figure 72: **Activation-level analysis to understand how different model pathways propagate memorization and generalization.** We observe that at downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ depicting that both residual connections and main branch are propagating memorization and generalization. Conversely, in most non-downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \gg 1$, showing that residual connections are the primary carriers. Furthermore, in the last two non-downsampling blocks (Block IDs 49, 50) $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ confirming the observation made in Section 4.2 that residual connections and main branch in those blocks have a similar minimal impact on memorization and generalization (Caltech101-ResNet152).



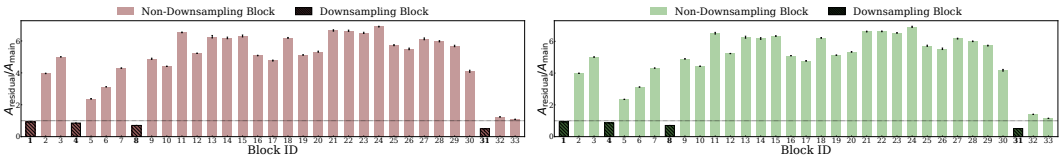
(a) Activation Analysis for Memorization (b) Activation Analysis for Generalization

Figure 73: **Activation-level analysis to understand how different model pathways propagate memorization and generalization.** We observe that at downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ depicting that both residual connections and main branch are propagating memorization and generalization. Conversely, in most non-downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \gg 1$, showing that residual connections are the primary carriers. Furthermore, in the last two non-downsampling blocks (Block IDs 32, 33) $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ confirming the observation made in Section 4.2 that residual connections and main branch in those blocks have a similar minimal impact on memorization and generalization (Caltech101-WideResnet101).



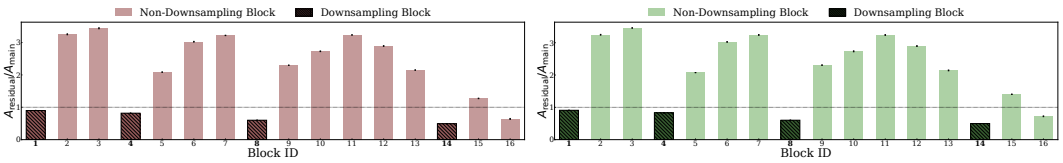
(a) Activation Analysis for Memorization (b) Activation Analysis for Generalization

Figure 74: **Activation-level analysis to understand how different model pathways propagate memorization and generalization.** We observe that at downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ depicting that both residual connections and main branch are propagating memorization and generalization. Conversely, in most non-downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \gg 1$, showing that residual connections are the primary carriers. Furthermore, in the last two non-downsampling blocks (Block IDs 32, 33) $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ confirming the observation made in Section 4.2 that residual connections and main branch in those blocks have a similar minimal impact on memorization and generalization (Oxford Flower-102-ResNeXt101).



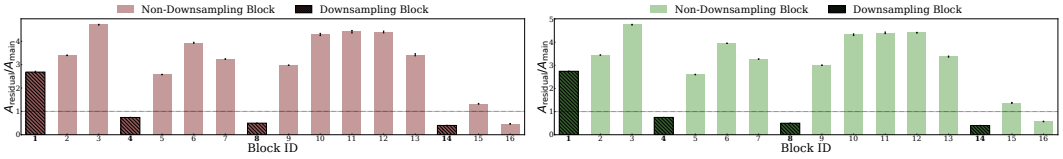
(a) Activation Analysis for Memorization (b) Activation Analysis for Generalization

Figure 75: **Activation-level analysis to understand how different model pathways propagate memorization and generalization.** We observe that at downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ depicting that both residual connections and main branch are propagating memorization and generalization. Conversely, in most non-downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \gg 1$, showing that residual connections are the primary carriers. Furthermore, in the last two non-downsampling blocks (Block IDs 32, 33) $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ confirming the observation made in Section 4.2 that residual connections and main branch in those blocks have a similar minimal impact on memorization and generalization (Places365-ResNet101).



(a) Activation Analysis for Memorization (b) Activation Analysis for Generalization

Figure 76: **Activation-level analysis to understand how different model pathways propagate memorization and generalization.** We observe that at downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ depicting that both residual connections and main branch are propagating memorization and generalization. Conversely, in most non-downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \gg 1$, showing that residual connections are the primary carriers. Furthermore, in the last two non-downsampling blocks (Block IDs 15, 16) $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ confirming the observation made in Section 4.2 that residual connections and main branch in those blocks have a similar minimal impact on memorization and generalization (Places365-ResNet50).



(a) Activation Analysis for Memorization (b) Activation Analysis for Generalization

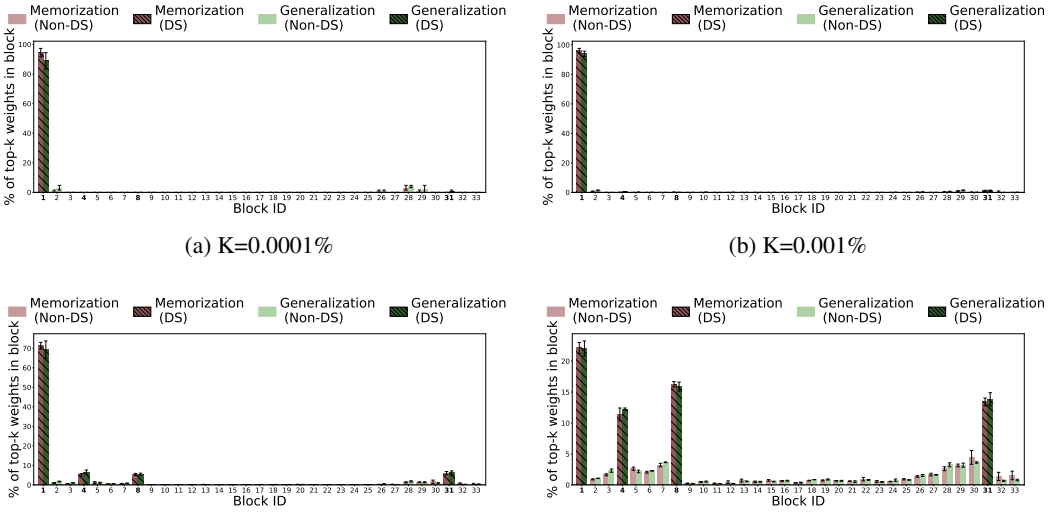
Figure 77: **Activation-level analysis to understand how different model pathways propagate memorization and generalization.** We observe that at downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ depicting that both residual connections and main branch are propagating memorization and generalization. Please note that in (Oxford-IIIT Pet-ResNet34) setup, $\frac{A_{\text{residual}}}{A_{\text{main}}} > 1$ for the first downsampling block, which probably means that the residual connection is slightly more dominant in the block. Conversely, in most non-downsampling blocks $\frac{A_{\text{residual}}}{A_{\text{main}}} \gg 1$, showing that residual connections are the primary carriers. Furthermore, in the last two non-downsampling blocks (Block IDs 15, 16) $\frac{A_{\text{residual}}}{A_{\text{main}}} \approx 1$ confirming the observation made in Section 4.2 that residual connections and main branch in those blocks have a similar minimal impact on memorization and generalization (Oxford-IIIT Pet-ResNet34).

C.6 Weights Contributing to memorization and Generalization are entangled and concentrated in Downsampling Blocks

In this section, we corroborate our findings that weights contributing to memorization and generalization are entangled and primarily present in downsampling block. We show this across different top-K weights ratios (0.0001%, 0.001%, 0.01%, and 0.1%), models (ResNet152, ResNet101, ResNet50, ResNet34, ResNeXt101, and WideResNet101), and datasets (Places365, CIFAR10, Oxford IIIT-Pet, and Oxford Flower-102) in Sections C.6.1 & C.6.2.

C.6.1 Memorization and Generalization Weights are Primarily Concentrated in Downsampling Blocks

From Figs. 78, 79, 80, 81, 82, 83, 84, 85, and 86, we observe that weights contributing to memorization and generalization are primarily concentrated in downsampling blocks across all top-K weights ratios, datasets and models.



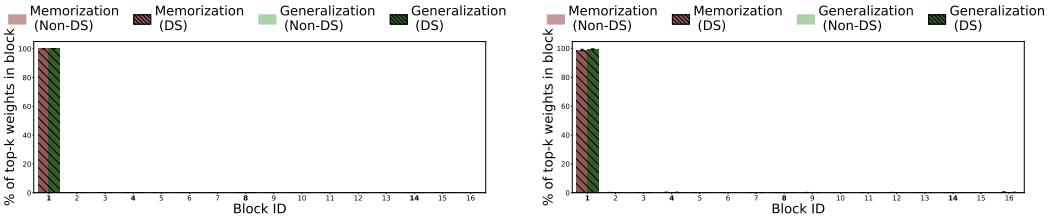
(a) K=0.0001%

(b) K=0.001%

(c) K=0.01%

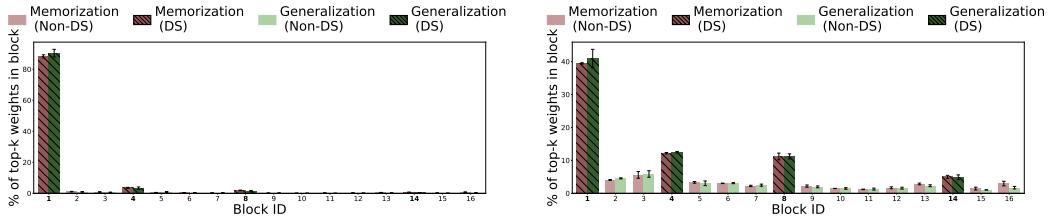
(d) K=0.1%

Figure 78: Memorization and Generalization weights are primarily concentrated in downsampling blocks across all tested K ratios (Places365-ResNet101, DS: Downsampling, NDS: Non-Downsampling).



(a) K=0.0001%

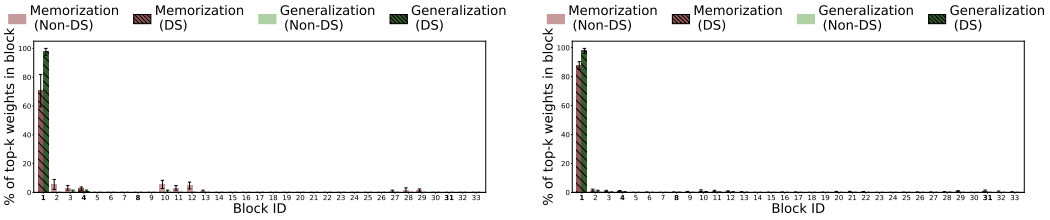
(b) K=0.001%



(c) K=0.01%

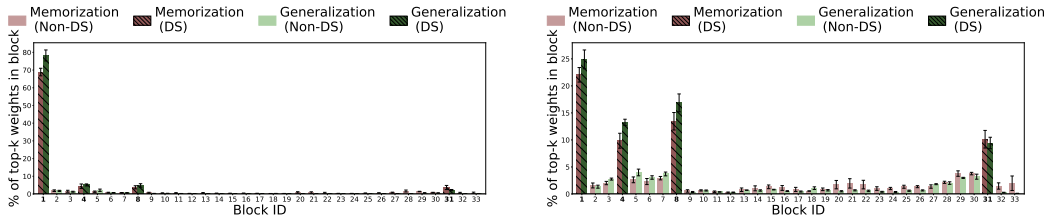
(d) K=0.1%

Figure 79: Memorization and Generalization weights are primarily concentrated in downsampling blocks across all tested K ratios (Places365-ResNet50, DS: Downsampling, NDS: Non-Downsampling).



(a) K=0.0001%

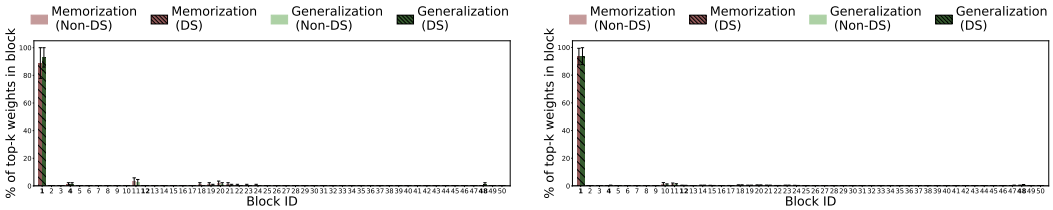
(b) K=0.001%



(c) K=0.01%

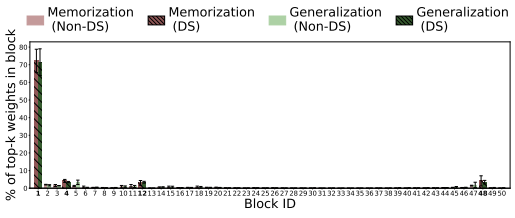
(d) K=0.1%

Figure 80: Memorization and Generalization weights are primarily concentrated in downsampling blocks across all tested K ratios (Oxford Flower-102-ResNet101, DS: Downsampling, NDS: Non-Downsampling).

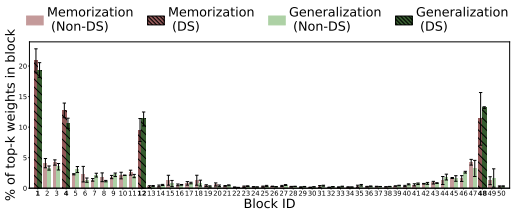


(a) K=0.0001%

(b) K=0.001%

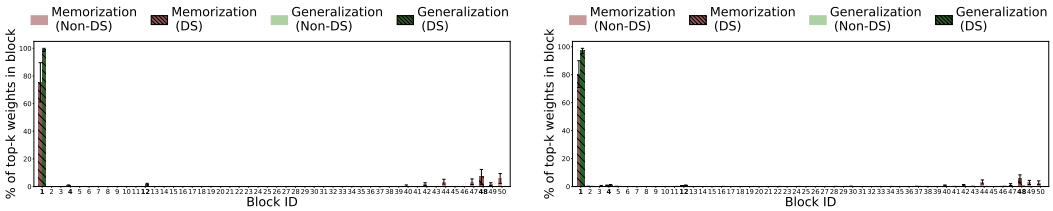


(c) K=0.01%



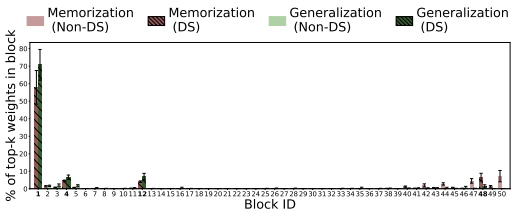
(d) K=0.1%

Figure 81: Memorization and Generalization weights are primarily concentrated in downsampling blocks across all tested K ratios (CIFAR10-ResNet152, DS: Downsampling, NDS: Non-Downsampling).

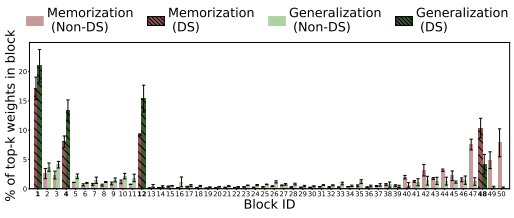


(a) K=0.0001%

(b) K=0.001%

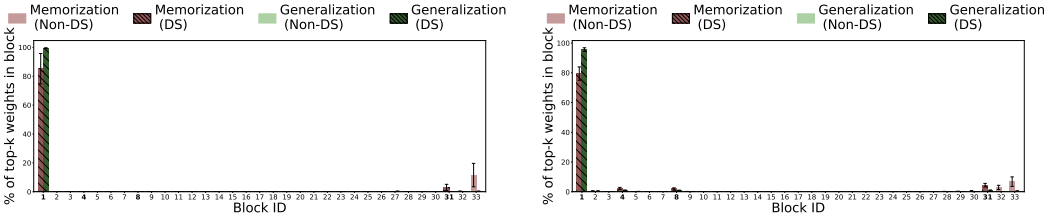


(c) K=0.01%



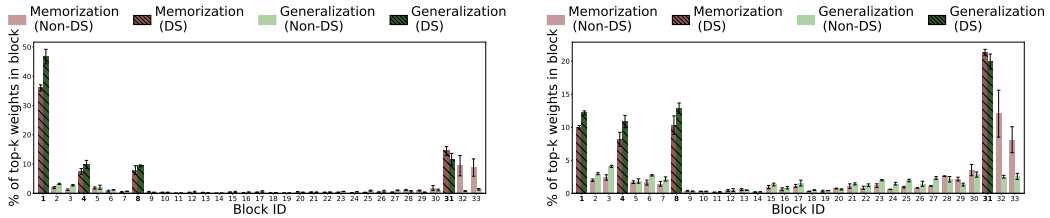
(d) K=0.1%

Figure 82: Memorization and Generalization weights are primarily concentrated in downsampling blocks across all tested K ratios (Caltech101-ResNet152, DS: Downsampling, NDS: Non-Downsampling).



(a) K=0.0001%

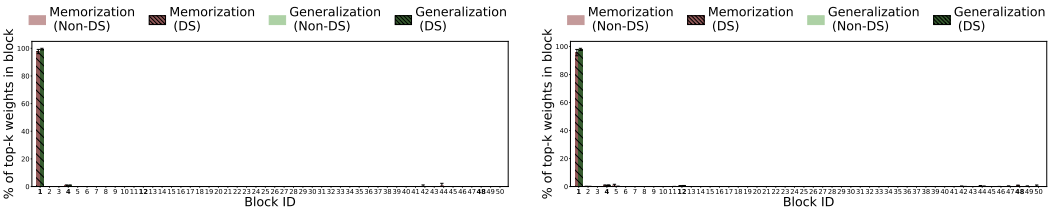
(b) K=0.001%



(c) K=0.01%

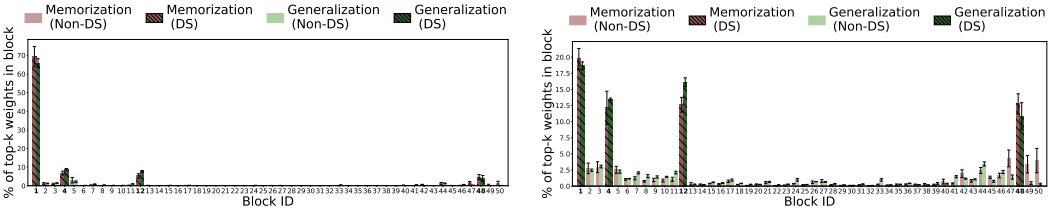
(d) K=0.1%

Figure 83: Memorization and Generalization weights are primarily concentrated in downsampling blocks across all tested K ratios (Caltech101-WideResNet101, DS: Downsampling, NDS: Non-Downsampling).



(a) K=0.0001%

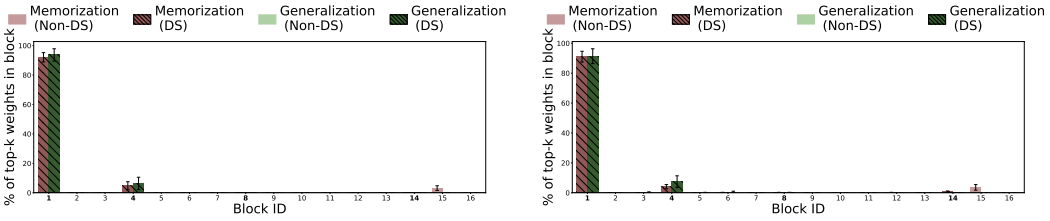
(b) K=0.001%



(c) K=0.01%

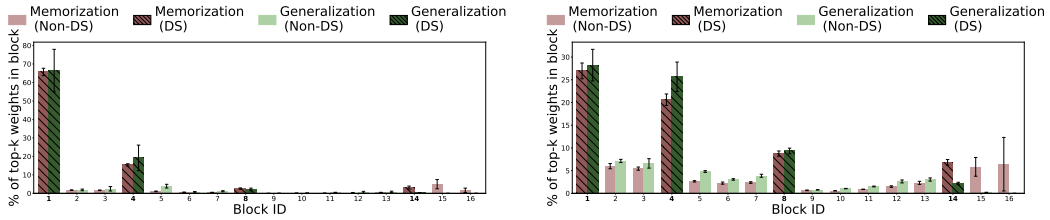
(d) K=0.1%

Figure 84: Memorization and Generalization weights are primarily concentrated in downsampling blocks across all tested K ratios (Oxford IIIT-Pet-ResNet152, DS: Downsampling, NDS: Non-Downsampling).



(a) K=0.0001%

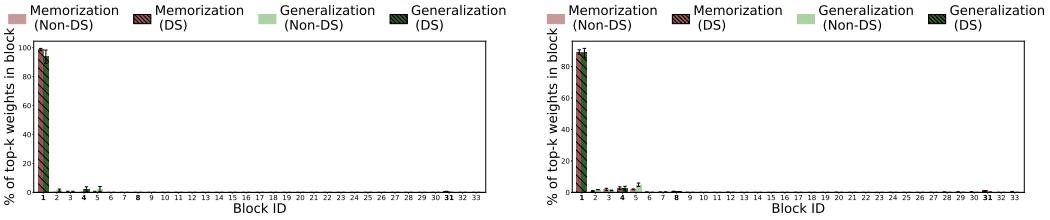
(b) K=0.001%



(c) K=0.01%

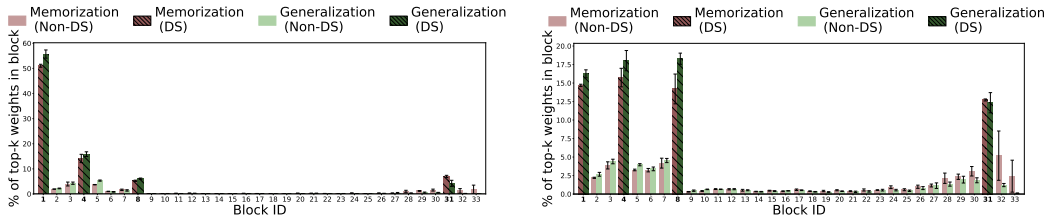
(d) K=0.1%

Figure 85: Memorization and Generalization weights are primarily concentrated in downsampling blocks across all tested K ratios (Oxford IIIT-Pet-ResNet34, DS: Downsampling, NDS: Non-Downsampling).



(a) K=0.0001%

(b) K=0.001%



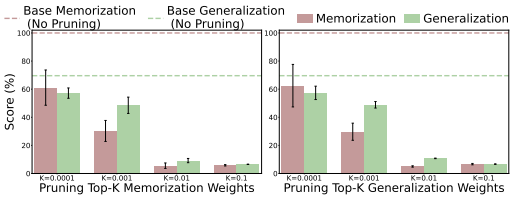
(c) K=0.01%

(d) K=0.1%

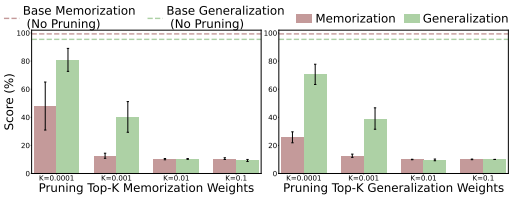
Figure 86: Memorization and Generalization weights are primarily concentrated in downsampling blocks across all tested K ratios (Oxford IIIT-Pet-ResNeXt101, DS: Downsampling, NDS: Non-Downsampling).

C.6.2 Weights Contributing to Memorization and Generalization are Entangled

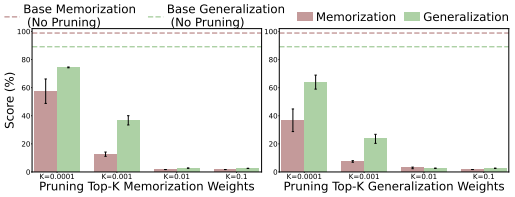
From Fig. 87, we can clearly observe that across all models and datasets, weights contributing to memorization and generalization are entangled, where pruning/removal of either one of the set of weights causes a drop in both memorization and generalization.



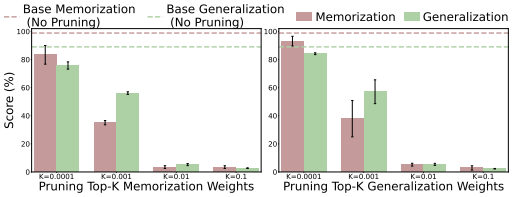
(a) Places365 – ResNet50



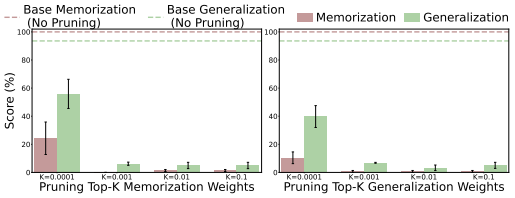
(b) CIFAR-10 – ResNet152



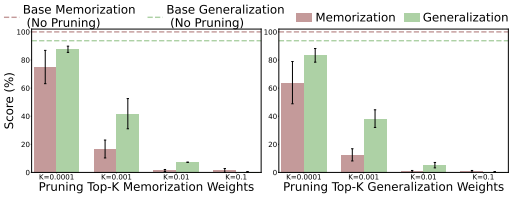
(c) Oxford-IIIT Pet – ResNet152



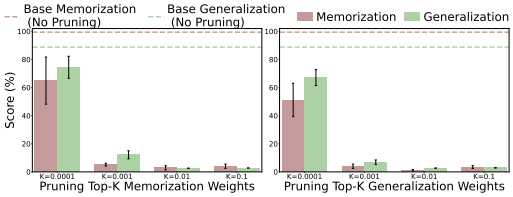
(d) Oxford-IIIT Pet – ResNet34



(e) Caltech101 – WideResNet101



(f) Caltech101 – ResNet152

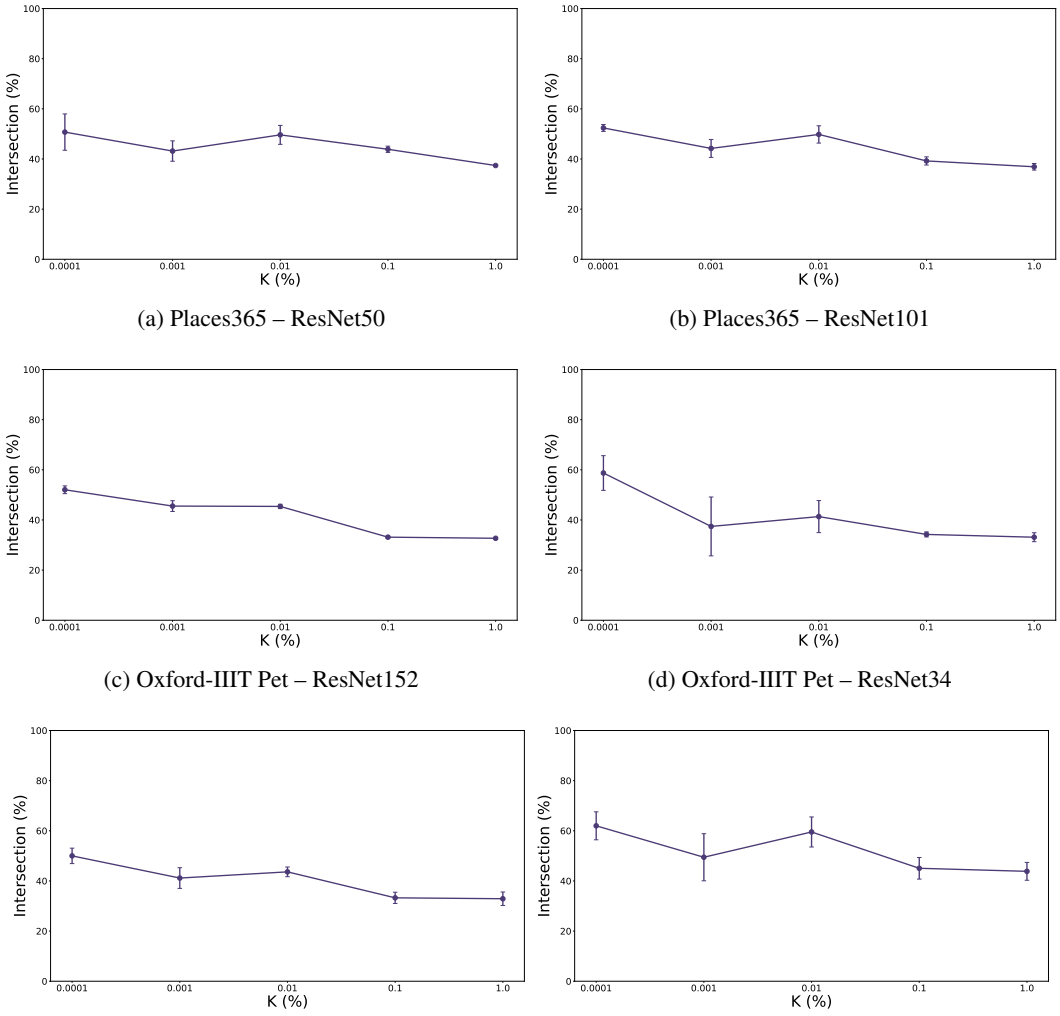


(g) Oxford-IIIT Pet – ResNeXt101

Figure 87: Weights contributing to Memorization and Generalization are entangled.

C.6.3 Substantial Overlap Between Memorization and Generalization Weights

In this section, we further show that memorization and generalization weights are overlapped, further showing evidence regarding their entanglement. The results are presented in Fig. 88



(a) Places365 – ResNet50

(b) Places365 – ResNet101

(c) Oxford-IIIT Pet – ResNet152

(d) Oxford-IIIT Pet – ResNet34

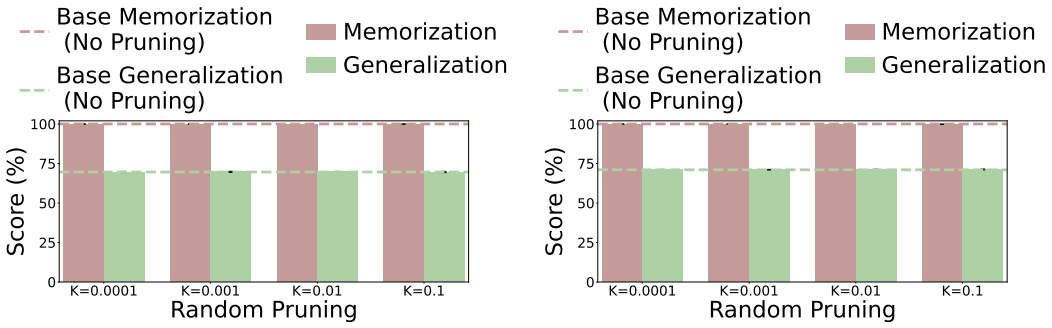
(e) Oxford-IIIT Pet – ResNeXt101

(f) CIFAR-10 – ResNet152

Figure 88: Memorization and generalization weight sets show overlap across all K ratios, further providing evidence on their entanglement.

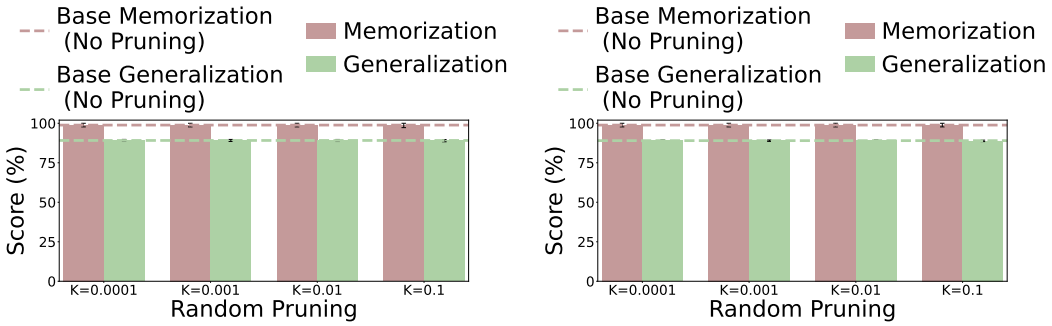
C.6.4 Baseline Comparison Against Random Pruning

In this section, we want to show that pruning a random, size-matched set of weights has negligible impact on memorization and generalization compared to pruning the top-K memorization and generalization identified weights, confirming that the observed degradation is not merely a consequence of removing network capacity but is specifically attributable to the identified weights. The results are presented in Fig. 89.



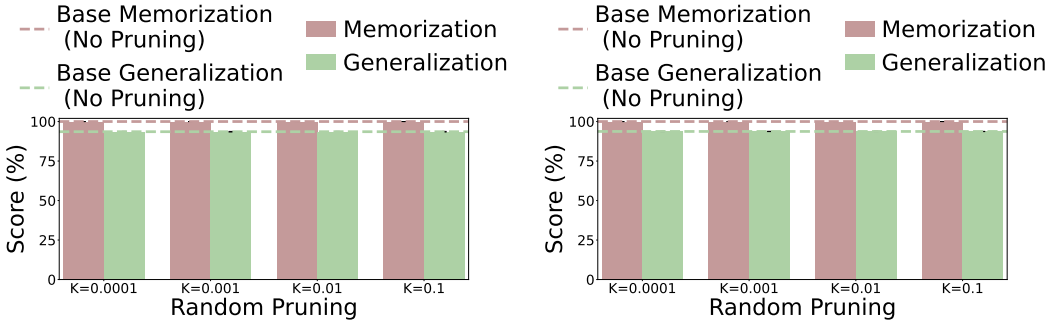
(a) Places365 – ResNet50

(b) Places365 – ResNet101



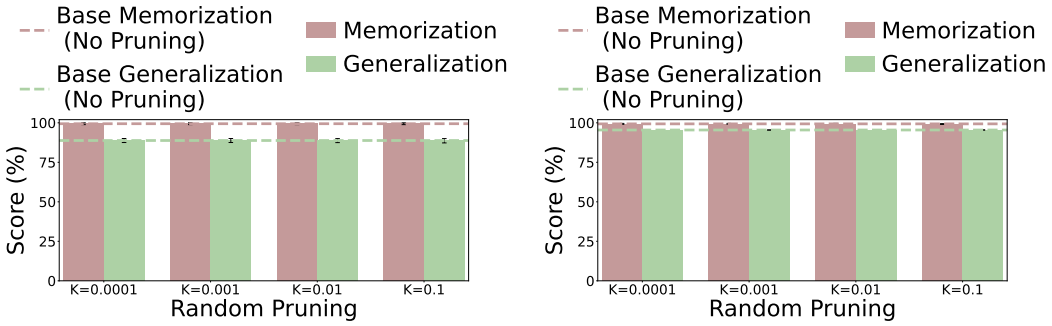
(c) Oxford-IIIT Pet – ResNet152

(d) Oxford-IIIT Pet – ResNet34



(e) Caltech101 – WideResNet101

(f) Caltech101 – ResNet152



(g) Oxford-IIIT Pet – ResNeXt101

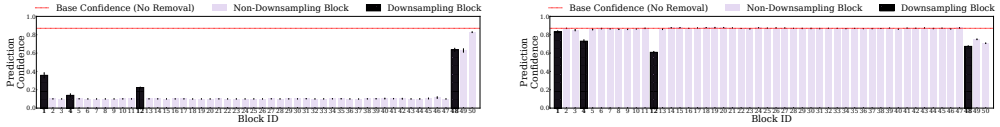
(h) CIFAR-10 – ResNet152

Figure 89: Random Pruning has almost no impact on memorization and generalization (This was done for 10 random seeds).

C.7 Model Confidence Prediction Analysis Across Blocks

In this section, we analyze how ablating residual connections and the main branch across different blocks influences the model's prediction confidence on the test set. These results further elucidate the complementary roles of downsampling and non-downsampling blocks, and of residual connections and the main branch, in shaping not only accuracy but also the confidence of the model's predictions. We also provide an interesting observation where both the residual connection and main branch in the last two non-downsampling blocks are responsible for ***confidence boosting*** of model's predictions, even though their ablation has negligible impact on test accuracy. The results corresponding to multiple datasets—CIFAR10, Places365, Oxford Flower-102, Oxford-IIIT Pet, and Caltech101—and architectures—ResNet34, ResNet50, ResNet101, ResNet152, WideResNet101, and ResNeXt101—are presented in Figs. 8, 90, 91, 92, 93, 94, 95, 96, 97, 98 for 2% label noise, Figs. 99, 100, 101, 102, 103, 104, 105, 106, 107, 108 for 5% label noise, and Figs. 109, 110, 111, 112, 113, 114, 115, 116, 117, 118 for 10% label noise,

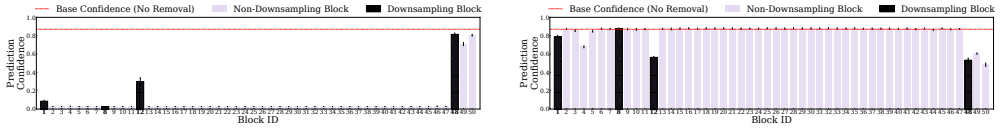
C.7.1 2% Label Noise



(a) Impact of Residual Connections on Prediction Confidence across blocks (CIFAR10-ResNet152)

(b) Impact of Main Branch on Prediction Confidence across blocks (CIFAR10-ResNet152)

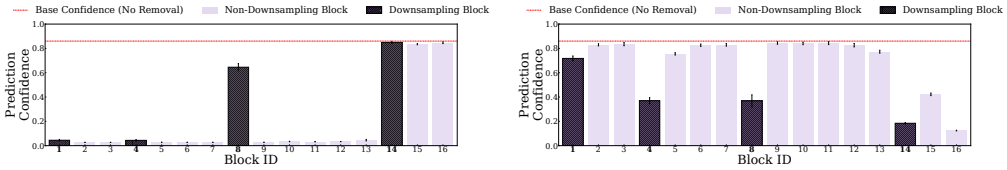
Figure 90: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 49, 50) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet152)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet152)

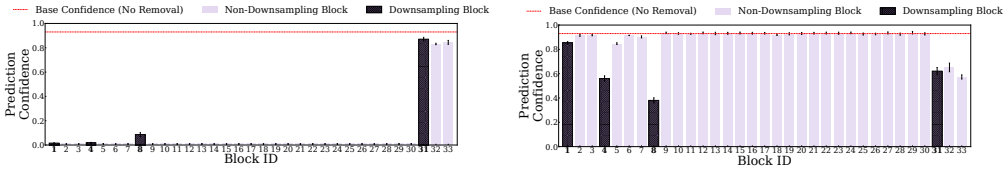
Figure 91: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 49, 50) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet34)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet34)

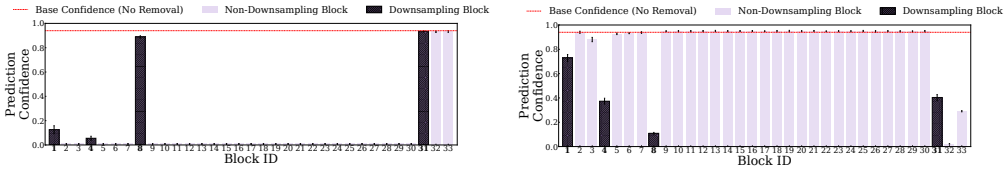
Figure 92: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 15, 16) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford Flower-102-ResNet101)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford Flower-102-ResNet101)

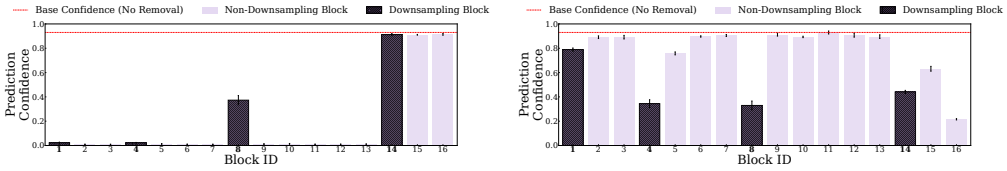
Figure 93: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford Flower-102-ResNeXt101)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford Flower-102-ResNeXt101)

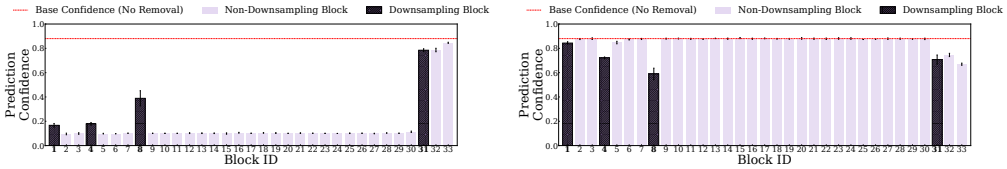
Figure 94: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford Flower-102-ResNet50)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford Flower-102-ResNet50)

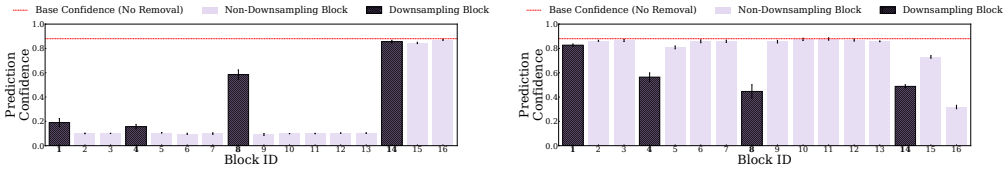
Figure 95: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 15, 16) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Places365-ResNet101)

(b) Impact of Main Branch on Prediction Confidence across blocks (Places365-ResNet101)

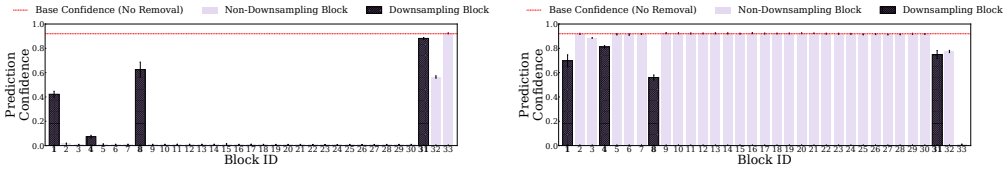
Figure 96: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Places365-ResNet50)

(b) Impact of Main Branch on Prediction Confidence across blocks (Places365-ResNet50)

Figure 97: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blockss (Block IDs 15, 16) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)

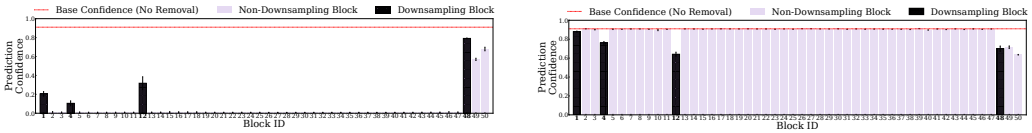


(a) Impact of Residual Connections on Prediction Confidence across blocks (Caltech101-WideResNet101)

(b) Impact of Main Branch on Prediction Confidence across blocks (Caltech101-WideResNet101)

Figure 98: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)

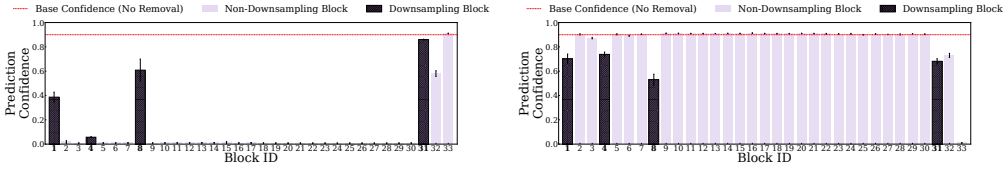
C.7.2 5% LABEL NOISE



(a) Impact of Residual Connections on Prediction Confidence across blocks (Caltech101-ResNet152, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Caltech101-ResNet152, 5% Label Noise)

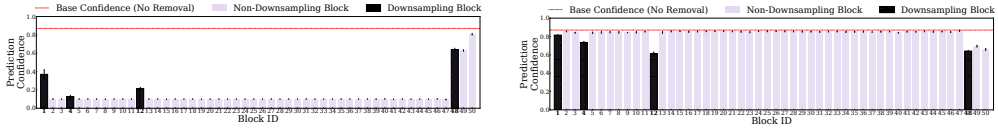
Figure 99: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 49, 50) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. Consistent trends observed for other architectures and datasets in Appendix C.7. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Caltech101-WideResNet101, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Caltech101-WideResNet101, 5% Label Noise)

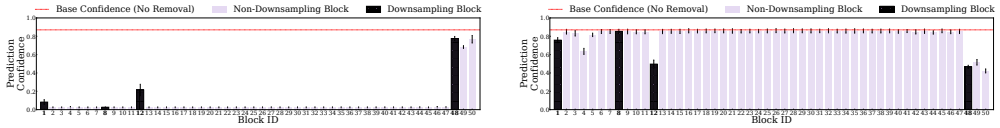
Figure 100: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (CIFAR10-ResNet152, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (CIFAR10-ResNet152, 5% Label Noise)

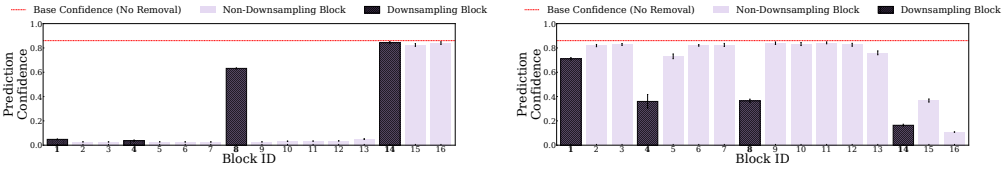
Figure 101: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 49, 50) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet152, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet152, 5% Label Noise)

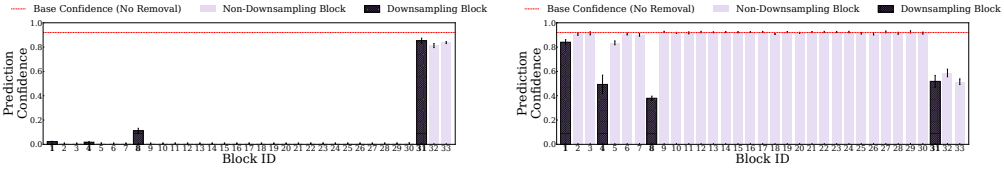
Figure 102: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 49, 50) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet34, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet34, 5% Label Noise)

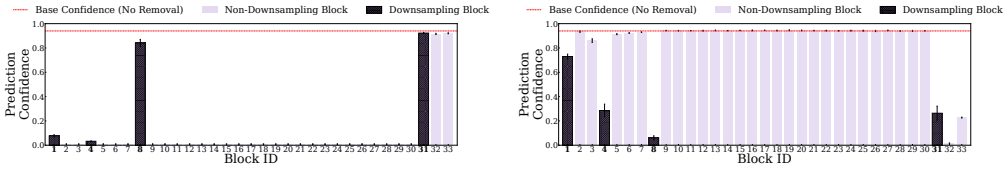
Figure 103: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 15, 16) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford Flower-102-ResNet101, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford Flower-102-ResNet101, 5% Label Noise)

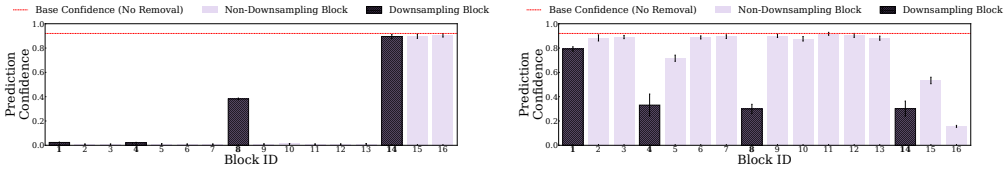
Figure 104: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford Flower-102-ResNeXt101, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford Flower-102-ResNeXt101, 5% Label Noise)

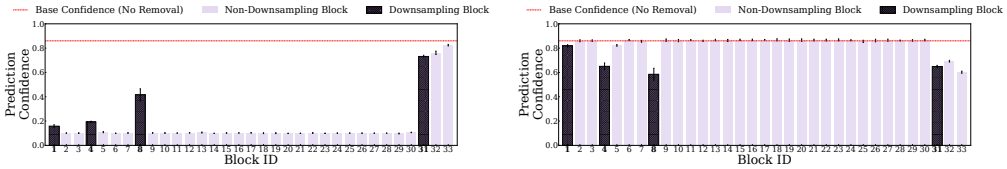
Figure 105: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford Flower-102-ResNet50, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford Flower-102-ResNet50, 5% Label Noise)

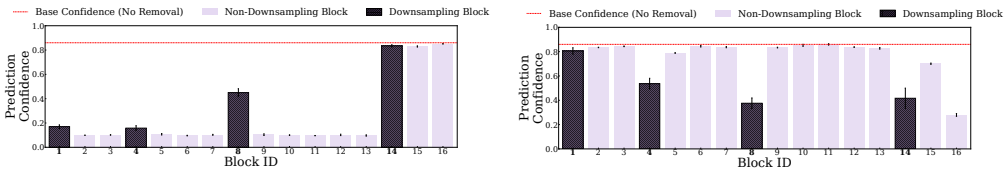
Figure 106: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 15, 16) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Places365-ResNet101, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Places365-ResNet101, 5% Label Noise)

Figure 107: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)

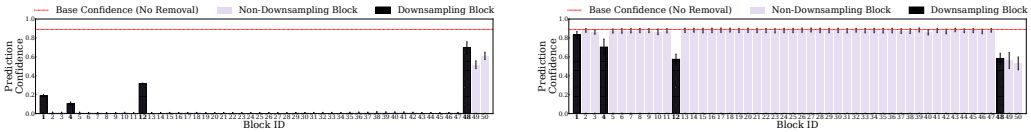


(a) Impact of Residual Connections on Prediction Confidence across blocks (Places365-ResNet50, 5% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Places365-ResNet50, 5% Label Noise)

Figure 108: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 15, 16) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)

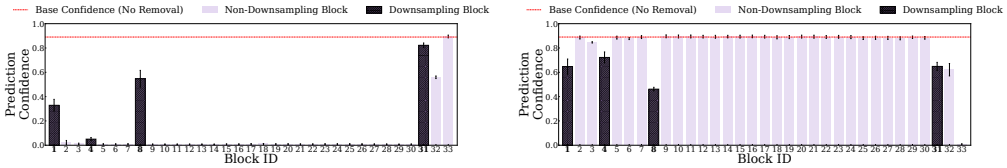
C.7.3 10% LABEL NOISE



(a) Impact of Residual Connections on Prediction Confidence across blocks (Caltech101-ResNet152, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Caltech101-ResNet152, 10% Label Noise)

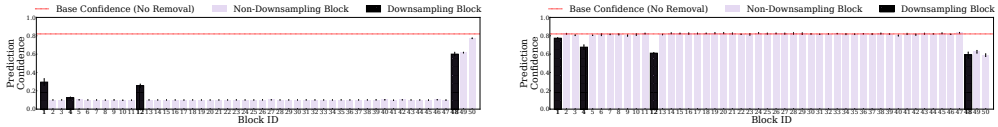
Figure 109: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 49, 50) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. Consistent trends observed for other architectures and datasets in Appendix C.7. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Caltech101-WideResNet101, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Caltech101-WideResNet101, 10% Label Noise)

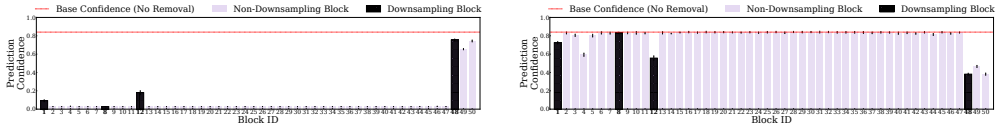
Figure 110: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (CIFAR10-ResNet152, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (CIFAR10-ResNet152, 10% Label Noise)

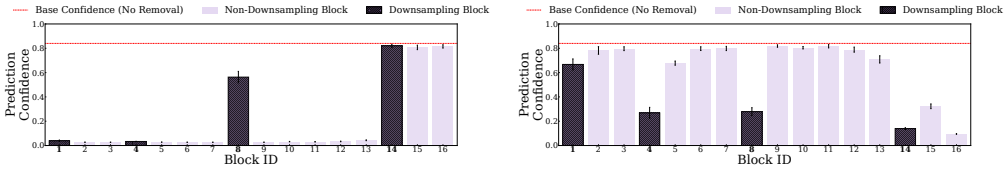
Figure 111: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 49, 50) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet152, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet152, 10% Label Noise)

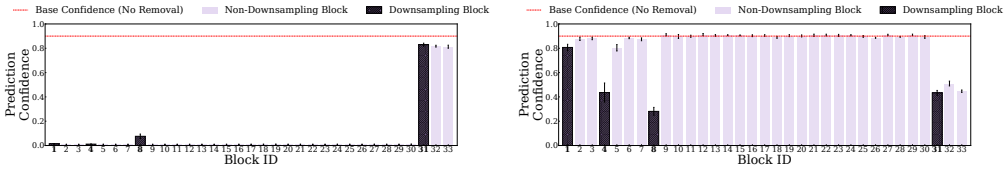
Figure 112: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 49, 50) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet34, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford-IIIT-Pet-ResNet34, 10% Label Noise)

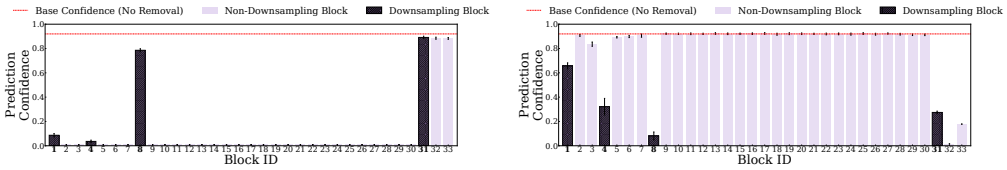
Figure 113: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 15, 16) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford Flower-102-ResNet101, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford Flower-102-ResNet101, 10% Label Noise)

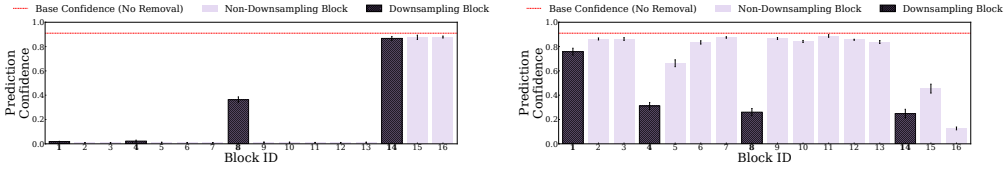
Figure 114: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford Flower-102-ResNeXt101, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford Flower-102-ResNeXt101, 10% Label Noise)

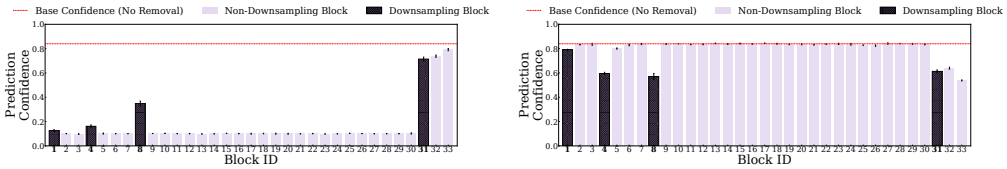
Figure 115: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Oxford Flower-102-ResNet50, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Oxford Flower-102-ResNet50, 10% Label Noise)

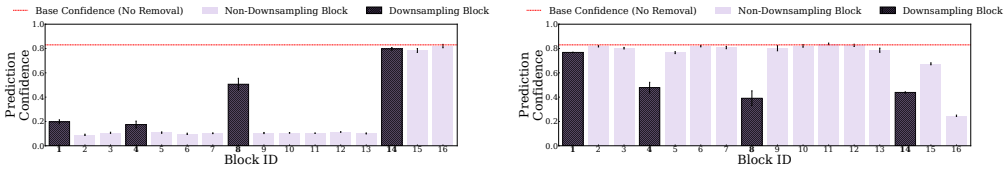
Figure 116: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 15, 16) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Places365-ResNet101, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Places365-ResNet101, 10% Label Noise)

Figure 117: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 32, 33) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)



(a) Impact of Residual Connections on Prediction Confidence across blocks (Places365-ResNet50, 10% Label Noise)

(b) Impact of Main Branch on Prediction Confidence across blocks (Places365-ResNet50, 10% Label Noise)

Figure 118: **Impact of residual connections and main branch ablations on prediction confidence.** Ablating residual connections (a) leads to a significant drop in prediction confidence for most non-downsampling blocks than downsampling ones. In contrast, removing non-downsampling main branch (b) has a limited influence in most cases, while removing downsampling main branch blocks shows a more noticeable drop. Interestingly, ablating the residual connections & main branch in the last two non-downsampling blocks (Block IDs 15, 16) slightly reduces prediction confidence despite negligible changes in test accuracy, suggesting a subtle role of these blocks in *confidence boosting*. (Results are averaged across the whole test set for 3 seeds, Zero Ablation)

D Limitations of Our Work

Our study relies on noisy labels (a commonly adopted approach) to study memorization. Alternative strategies—such as identifying memorized examples via forgetting dynamics or influence-based methods—may provide complementary insights but were not considered due to their additional computational overhead. Finally, as stated in the title, this work is restricted to ResNet-style architectures. Although other architectures like transformers also employ residual connections, they do not include downsampling blocks analogous to those in ResNets, making a direct comparison non-trivial (i.e, out of scope of this paper) and hence requires a separate, dedicated analysis.