Feature Visualization of Deep Neural Networks

Engin Deniz Erkan
Department of Data Informatics
Middle East Technical University
Ankara, TURKEY
engin.erkan@metu.edu.tr

***Abstract*— This paper explores visualization techniques to unveil the operational dynamics of deep neural networks, focusing primarily on feature visualization—a crucial method for understanding the complex representations within these models. The study employs a generic algorithm to comprehensively investigate feature visualization in image classifiers, implementing it across two prominent architectures: ResNet and Vision Transformer (VIT). The impact of diverse augmentation strategies is scrutinized, encompassing six augmentations for both ResNet and VIT alongside a single augmentation for comparative analysis. The findings contribute valuable insights into the detailed relationships between model architecture, augmentation methods, and the interpretability of learned features.**

Keywords— Feature Visualization, Deep Neural Networks, Visualization Techniques, Activation Maximization, Deep Dream, Style Transfer

I. Introduction

In deep neural networks, feature visualization is a method that is frequently used to understand and display the features that the network has learned. It is an essential tool for understanding and clarifying the complex representations these models learn. By using this method, researchers can visualize the internal workings of the network and gain insights into the features it has learned. Feature visualization is typically used to examine filters or features at different network layers, and it has shown to be especially useful in architectures intended for processing visual data, like Convolutional Neural Networks (CNNs). In order to explore and compare the visualization results across various models, this paper implements a generic algorithm that analyzes the field of feature visualization for image classifiers. Two well-known architectures, ResNet and Vision Transformer (VIT), are examined using different augmentation techniques in order to evaluate their feature visualization capabilities.

II. Literature Review

Natural signals frequently have hierarchical compositions, combining lower-level components to create higher-level features. Deep neural networks (DNNs) make use of this property. For example, local edge groups in photos establish patterns that further aggregate into discrete components, eventually forming recognized objects [1].

Feature visualization allows the investigation of the interactions between many neurons in addition to single neurons. The process of gradually changing the emphasis from one neuron to the other while investigating two neurons is similar to interpolating in generative models. However, the random character of the display presents a problem: dissimilar layouts prevent aligned comparisons. Without alignment, differences in objectives get overshadowed by significant layout disparities [2].

In feature visualization, examples are created iteratively to identify a neural network's emphasis points. The differentiability of the network with inputs reveals certain behaviors, but optimization problems still exist. The goals of optimization range from producing instances of output classes to comprehending particular characteristics, all of which call for sophisticated methods. Deeper insights are provided by style transfer and model inversion aims. Optimization allows for flexible investigations into combined neuron information encoding and feature development while also differentiating causality from correlation. Diversifying visualizations, comprehending neuronal connections, and minimizing artifacts throughout this process are complex tasks [3].

DNNs acquire discriminative traits from images, a characteristic shared with evolutionary processes. Yet, whether various DNNs capture identical attributes for each class remains to be determined. Notably, when images optimized for one DNN were utilized and transferred to another, the results highlighted intriguing patterns. While some images were specifically tailored for a single DNN, the overall traits of the networks played a role in their performance, as observed in evaluations on both MNIST and ImageNet DNNs. Experiments unveiled unexpected outcomes, demonstrating a diverse spectrum of images for each class rather than homogenous ones. Surprisingly, recognizable images still needed to evolve despite previous success with CPPN encoding. The primary aim was not to create unrecognizable images but rather to generate recognizable ones, defying the expected outcomes of either failure to achieve high confidence scores or creating unrecognizable images with low confidence. This divergence from predictions suggests a potential distinction between discriminative and generative models, leading to the production of highly confident yet unrecognizable images. The susceptibility of DNNs to deception is evident, as they assign high confidence to unrecognizable images. Furthermore, both gradient ascent and evolutionary algorithms yield two distinct types of "fooling images," highlighting the disparities between DNNs and human object recognition. These findings prompt questions about the generalization abilities and susceptibility of DNNs to sophisticated manipulations [4].

At the OpenVis Conference, the presenter discusses a novel approach, termed feature inversion or visualization, aimed at understanding complex neural network behavior beyond traditional data visualization methods. This method involves manipulating random noise to stimulate specific

features within the network, generating images that capture what these features respond to most. By exploring one particular layer's feature inversions, the speaker reveals how these representations align with human-understandable concepts like dogs, bicycle wheels, cars, and actions. Further inspection through various network layers unveils a progression from basic patterns to more complex object concepts, illuminating the network's hierarchical comprehension of images. Moreover, the speaker emphasizes the network's utilization of combinations of these features and their linear interpolations, showcasing the network's ability to comprehend images through a rich understanding developed layer by layer. This technique not only offers insights into how neural networks process information but also highlights their ability to represent and understand complex visual features in a manner interpretable to humans [5].

A method for creating images in deep neural networks called excerpt elucidates works by optimizing input pixels as opposed to network weights. Traditional gradient ascent struggles with local maximums and frequently produces noise. In order to get around this, a technique called gradient ascent gradually presents low-frequency information before high-frequency details, which results in workable visualizations. Both gradient and Gaussian blur on the picture are recommended for this iterative optimization; the former yields superior outcomes. Deeper networks, such as GoogLeNet, present challenges because decreasing blurring can result in high-frequency noise saturation, potentially because of gradient instability linked to network depth. The use of GoogLeNet's auxiliary classifiers for image generation demonstrates how they can generate images with different levels of noise and structure depending on how close they are to early network layers. Combining gradients from these classifiers in sequence improves structure and high-frequency detail retention. Additionally, scaling up images at intervals enhances image quality and size for optimization, akin to Deepdream's "octaves" methodology, overcoming constraints imposed by limited input dimensions [6].

The study conducted by Fong and Vedaldi introduces a novel framework designed to elucidate machine learning model predictions, particularly in critical domains like medical diagnosis and autonomous systems. While existing image saliency methods reveal neural network focus areas in images, limitations persist. This newly proposed framework offers a versatile solution capable of acquiring diverse explanations for any black box algorithm. Focused on identifying pivotal image segments influencing classifier decisions, this model-agnostic approach enables interpretable image perturbations, allowing comprehensive testing. The work culminates in proposing a formalized meta-predictor framework for learning explanations. This new image saliency paradigm significantly enhances interpretability by pinpointing critical image components, facilitating understanding, and exposing vulnerabilities within neural networks [7].

The live activation visualizer offers real-time insight into a trained convnet's processing of images or videos, fostering a clearer understanding of the network's inner workings. Additionally, the feature visualizer utilizes regularization methods, enhancing the interpretability of each layer's features within a DNN. Notably, both tools are open-source and boast user-friendly setups. The conducted by Yosinski and Fuchs underscores the significance of these tools in interpreting trained neural networks, potentially advancing methodologies, and future research endeavors. It also explores intriguing possibilities, such as an alternate world decomposition perception from that of humans and the crucial role of sparse connectivity in transfer learning. Introducing novel regularizations aims to augment model understanding and performance. Furthermore, it highlights the challenge of creating generative models capable of producing realistic images, suggesting the utilization of discriminative parameters in this pursuit [8].

The conducted by Erhan et al. examines the qualitative interpretations of high-level features within deep architectures, particularly in vision datasets, aiming to offer insights into the inner workings of these models. Comparing techniques applied on Stacked Denoising Autoencoders and Deep Belief Networks, the study reveals that interpretation at the unit level is feasible, straightforward, and consistent across these methods. One notable approach explored is "activation maximization," aimed at finding input patterns that maximize the activation of hidden units in a neural network. This technique, facilitated by gradient ascent optimization, demonstrates applicability to networks where gradients are computable. The comparison of three techniques—activation maximization, sampling from a unit, and linear combination—across datasets confirms that higher-layer units encode more intricate features, showcasing differing learned features between deep architectures. Furthermore, the study suggests potential applications of these visualization tools in comparing network-learned features and exploring broader datasets and models. Future directions involve scrutinizing higher-level unit behavior in deep networks against features encoded by the visual cortex [9].

The study conducted by Guo, Luan, and Li investigates deep learning object detection algorithms, specifically classifying them into two types: two-stage and single-stage, with a focus on single-object detection models. A tailored feature visualization system for convolutional neural networks (CNNs) and object detection models is introduced to enhance observation and evaluation. In the evaluation, YOLOv3SE demonstrates reduced accuracy compared to YOLOv3 but excels in detecting smaller objects. Conversely, YOLOv3SPP maintains accuracy levels similar to YOLOv3, which is attributed to its effective reduction of redundant feature extraction. Modifications, such as adjusting network width and depth in YOLOv3m and YOLOv3s models, lead to compromised accuracy but substantially improve detection speed and training efficiency. These adapted models, YOLOv3m and YOLOv3s, prove advantageous in scenarios prioritizing swift detection over absolute precision [10].

III. Feature visualization

A. Residual Networks (ResNet)

ResNet, or Residual Networks, represents a pivotal advancement in deep learning architecture. One of its main features is its ability to make training deep neural networks easier by adding residual connections to mitigate the vanishing gradient issue. Because of these connections, the network can learn residual mappings, which facilitates optimization and keeps accuracy from declining as network depth rises. The skip connections in ResNet facilitate direct information transfer between layers, which helps effectively learn shallow and deep features. However, when the model parameters are raised, the insertion of these skip links comes at a more considerable computational cost. Furthermore, the depth of ResNet may cause overfitting in situations where the dataset could be more extensive. As a whole, ResNet has demonstrated remarkable success in various computer vision tasks, particularly image classification, and remains a foundational architecture in the deep learning landscape.

B. Vision Transformer (VIT)

The Vision Transformer (VIT) stands out in the context of the project on feature visualization of deep neural networks. Its primary benefit is that it can capture complex long-range dependencies in images, which makes it especially good at comprehending everything completely. The ability to visualize complex and subtle patterns connected to the selected target classes becomes essential. Because VIT can scale to multiple input resolutions, it can effectively extract features that match the diverse properties of these target classes; however, it is essential to acknowledge that the transformer-based architecture of VIT might face challenges in scenarios where preserving spatial hierarchies is paramount, as in the case of object localization. Moreover, the computational burden of self-attention processes may affect the effectiveness of feature visualization, particularly when different augmentations for image classifiers are taken into account. Despite these drawbacks, VIT is a valuable model for feature visualization due to its competitive efficiency in capturing global dependencies.

C. Augmentations

Augmentations play a vital role in feature visualization, serving as a fundamental method to boost the robustness and generalization capabilities of deep neural networks. In the context of the project, augmentations are crucial for creating diverse input samples that enable the network to learn invariant features and achieve better generalization across different target classes. The training data is varied by a number of augmentations, including random scaled cropping, horizontal flips, rotations, perspective transformations, Gaussian blur, and random grayscale. This diversity encourages the model to recognize and emphasize the fundamental characteristics of the target classes during feature visualization. As a regularizing method, augmentations help avoid overfitting and encourage a more sophisticated comprehension of the unique traits connected to each class. To ensure that the model learns representations that are both reliable and representative of the underlying data distribution, the feature visualization process depends critically on the careful selection of augmentations.

D. Implimentation

This study explores feature visualization in deep neural networks, with a focus on ResNet and Vision Transformer (VIT) models. Four distinct models were examined, each employing different combinations of augmentations and architectures. Specifically, the investigation involved VIT with six augmentations, ResNet with six augmentations, ResNet with one augmentation, and VIT with one augmentation.

White Shark, Tarantula, Jellyfish, Zebra, Beer Bottle, Car Wheel, Electric Guitar, Sunglass, Tennis Ball, and Mushroom were the ten randomly chosen target classes for which feature visualizations were the focus of the investigation. For effective computing, Google Colab was utilized in conjunction with the T4 GPU.

The main focus of the research was using a custom implementation to train visualization models for ResNet and VIT. The training loop comprised 1000 iterations for each model and augmentation setting, employing a batch size of 8. The goal of the training process was to produce visuals that maximized the target class's activation. Each model and augmentation setting underwent corresponding optimization to enhance the model's capability to recognize and highlight features associated with the target class.

At regular intervals (every 100 iterations), the models were transitioned to evaluation mode, and images were generated based on the learned representations. The resulting visualizations offered insights into the changing feature representations for every model and augmentation approach and were then presented for comparative examination. The picture below shows how the image quality increases as the number of iterations increases.

Fig. 1. Feature visualization of a tennis ball created by ResNet

The findings advance the knowledge of the complex interactions between various models and augmentations that affect deep neural network interpretability. Contributing to the expanding body of knowledge in feature visualization and serving as a basis for additional research and development of

interpretability strategies in deep learning, the visualizations provide insights into learned features and their applicability to designated target classes.

The feature visualizations outputs for ten different classes and 1000 iterations are shown in the images below. To make comparison easier, images produced for the same class are presented side by side.

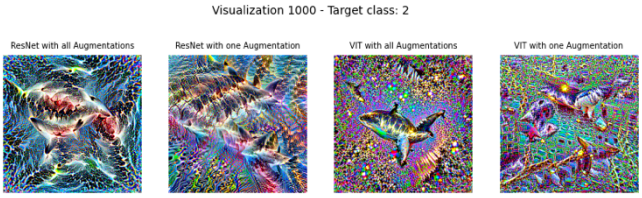

Fig. 2. Feature visualization of a white shark

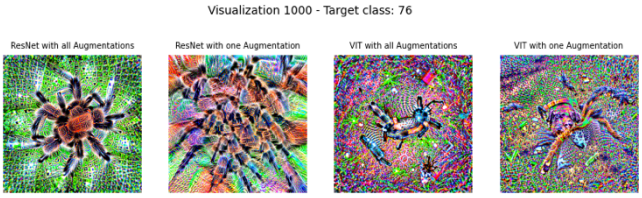

Fig. 3. Feature visualization of a tarantula

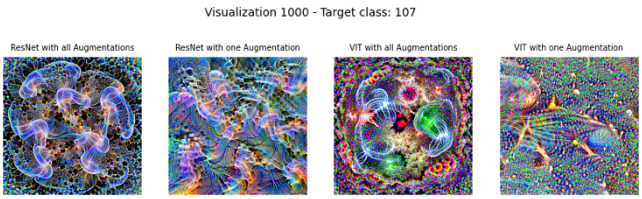

Fig. 4. Feature visualization of a jellyfish

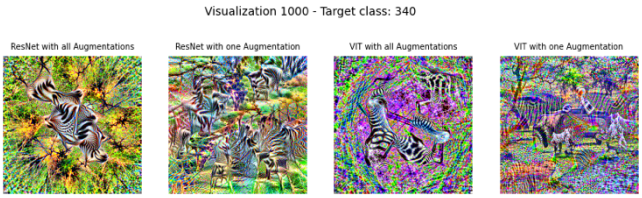

Fig. 5. Feature visualization of a zebra

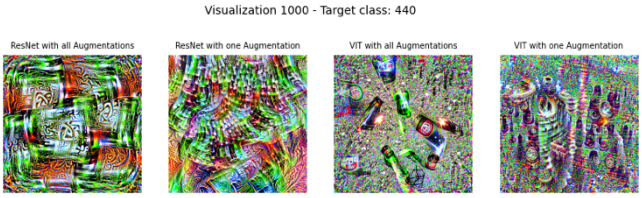

Fig. 6. Feature visualization of a beer bottle

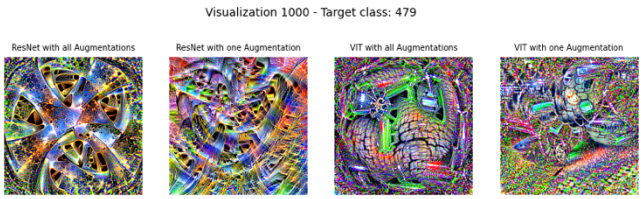

Fig. 7. Feature visualization of a car wheel

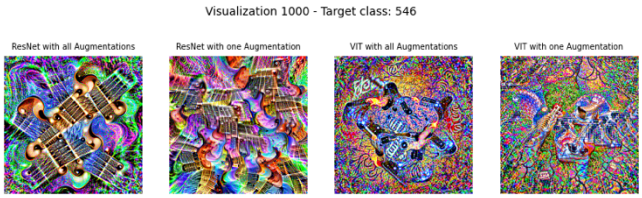

Fig. 8. Feature visualization of an electric guitar

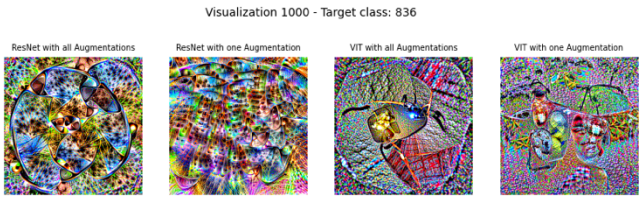

Fig. 9. Feature visualization of a sunglass

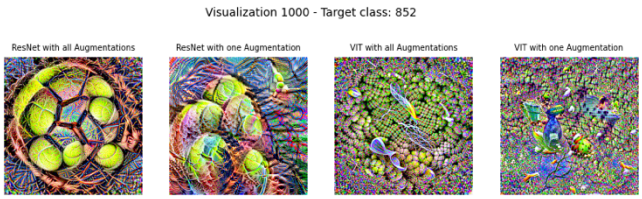

Fig. 10. Feature visualization of a tennis ball

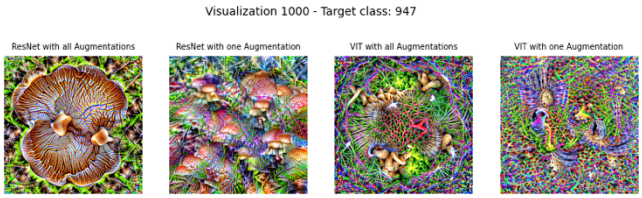

Fig. 11. Feature visualization of a mushroom

IV. Experiments

In the experimental phase, four distinct sets of images created for the same target class are used. These images were crafted through feature visualizations employing two distinct neural network architectures, ResNet and Vision Transformer (VIT), with different augmentation combinations. To assess the efficacy of the feature visualizations, a dedicated dataloader was implemented, facilitating the integration of PNG files with corresponding augmentations and target class labels. The dataset was further augmented, encompassing the reapplication of augmentation techniques to the previously generated images and the generation of new images. Each dataloader encapsulated 300 images, providing a substantial

dataset for experimental inspection. The experimental design involved subjecting the VIT model to predict the target class for both augmented and non-augmented sets of 300 ResNet images while reciprocally assessing the ResNet model's prediction of the target class for both augmented and non-augmented sets of 300 VIT images. These meticulously chosen experimental setups aimed to ascertain the robustness and generalizability of the feature visualizations across different neural network architectures and augmentation strategies. The figures below show some predictions made by ResNet and VIT models.

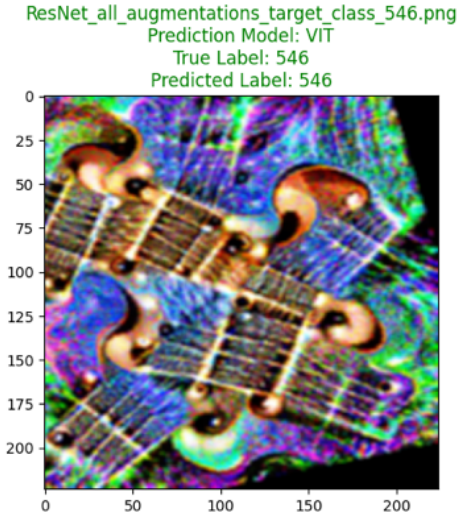

Fig. 12. Prediction made by VIT model of an electric guitar

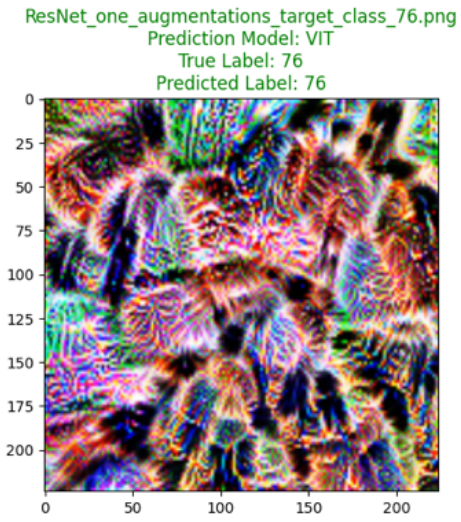

Fig. 13. Prediction made by VIT model of a tarantula

Figure 14 shows a prediction made by the ResNet model. Although the target class should have been white shark, the prediction was made as hammerhead shark.

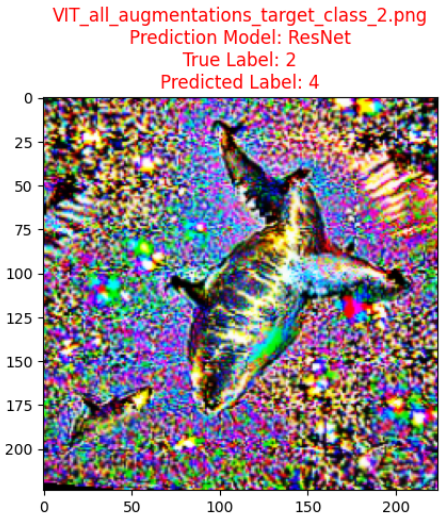

Fig. 14. Prediction made by ResNet model of a white shark

Figure 15 shows a prediction made by the ResNet model. Although the target class should have been zebra, the prediction was made as a T-shirt.

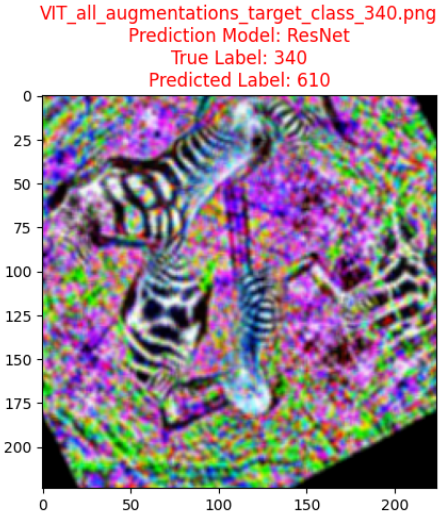

Fig. 15. Prediction made by ResNet model of a zebra

V. Experimental Results

The experimental results, which included a total of 1200 images, were examined and the table below was prepared.

TABLE I. Results

Prediction Model	Used Images	Accuracy
VIT	300 ResNet images (augmented)	60.67%
VIT	300 ResNet images (not augmented)	51.33%
ResNet	300 VIT images (augmented)	4.67%
ResNet	300 VIT images (not augmented)	2.33%

VI. Conclusion

In conclusion, the experimental outcomes reveal distinct patterns in the accuracy of predictions based on feature visualizations across different models and augmentation strategies. Firstly, the application of augmentations exhibits a noticeable impact on the predictive performance of the Vision Transformer (VIT) model. Notably, when subjected to feature visualizations of 300 ResNet images, the VIT model achieves a higher accuracy of 60.67% when augmented, compared to 51.33% without augmentation. This observation suggests a positive correlation between the augmentation complexity and the predictive capacity of the VIT model, with a tendency for more augmented images to better resemble the intended target class.

Secondly, a comparative analysis between ResNet and VIT models underscores substantial differences in the quality of generated images and subsequent prediction accuracies. Notably, the ResNet model exhibits low predictive capabilities, achieving accuracy levels of 4.67% and 2.33% for augmented and non-augmented sets of 300 VIT images, respectively. This result can be associated with the low quality of images created with the VIT model in terms of feature visualization. This marked discrepancy in accuracy suggests that images generated by ResNet more faithfully capture the features associated with the target class, emphasizing the effectiveness of ResNet in the context of feature visualization for image classification tasks.

In the broader context, these findings underscore the significance of meticulous experimentation and model selection in the feature visualization process. The observed correlations between augmentation complexity, model architecture, and prediction accuracy highlight the intricate interplay between these factors in shaping the efficacy of feature visualizations. As an essential component of deep neural network interpretability, feature visualization not only facilitates a better understanding of model behavior but also provides insights that can inform model refinement and development. Therefore, this study reaffirms the importance of feature visualization as a tool for elucidating the inner workings of complex neural networks, ultimately contributing to the advancement of interpretability and performance optimization in the field of deep learning.

VII. Future Works

In the process of feature visualization utilizing ResNet and VIT models, a comprehensive exploration was conducted through 1000 iterations to generate visual representations of learned features. Despite the substantial insights gained from this initial investigation, the achieved results were suboptimal when the generated images were subsequently employed in prediction models to ascertain the target class. It is hypothesized that extending the training duration beyond the limited scope of 1000 iterations could yield more refined and discriminative feature representations, potentially leading to enhanced performance in downstream tasks. Unfortunately, due to resource constraints, particularly regarding GPU availability, a more exhaustive exploration of feature visualization iterations could not be undertaken in the current study. This limitation prompts consideration for future investigations, where increased computational resources would enable a more thorough examination of the effects of prolonged training on feature visualization outcomes, thereby contributing to a more nuanced understanding of model interpretability and generalization capabilities. Figure 16 shows the effect of increasing the number of iterations in the training loop on feature visualization.

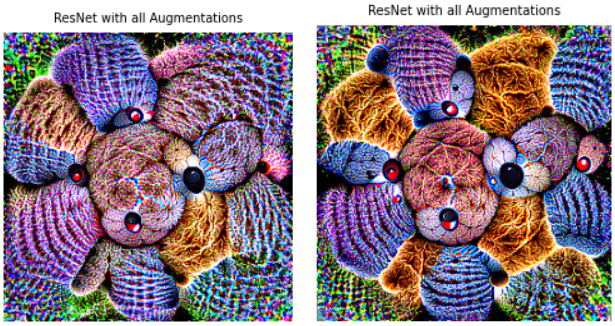

Fig. 16. Teddy bear feature visualization by 1000 iterations versus 10000 iterations

References

[1] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature 521, 436–444. doi: 10.1038/nature14539

[2] Mordvintsev, A., Pezzotti, N., Schubert, L., and Olah, C. Differentiable image parameterizations. Distill 3, 7 (2018), e12.

[3] Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. Distill 2, 11 (2017), e7.

[4] Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE conference on computer vision and pattern recognition (2015), pp. 427–436.

[5] Carter, S. (2018, July 31). Lessons from a year of distilling machine learning research [Video]. YouTube. https://www.youtube.com/watch?v=jlZsgUZaIyY

[6] Oygard, A. M. Visualizing googlenet classes (2015)

[7] R. C. Fong and A. Vedaldi, "Interpretable Explanations of Black Boxes by Meaningful Perturbation," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 3449-3457, doi: 10.1109/ICCV.2017.371.

[8] Yosinski, J., & Fuchs, T. (2015). Understanding Neural Networks Through Deep Visualization.

[9] Erhan, D.; Bengio, Y.; Courville, A.; Vincent, P. Visualizing higher-layer features of a deep network. Univ. Montr. 2009, 1341, 1.

[10] N. Guo, S. Luan and J. Li, "An Optimization Scheme of Object Detection Model Based on CNN Feature Visualization Method," *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, Xi'an, China, 2022, pp. 94-99, doi: 10.1109/ICIVC55077.2022.9887252.