Explaining Deep Image Classification Algorithms Using Grad-CAM

Prashant Bhardwaj
Saarland University
Saarland, Germany
prbh00001@stud.uni-saarland.de

Zurana Mehrin Ruhi
Saarland University
Saarland, Germany
zuru00001@stud.uni-saarland.de

Abstract

In this paper, we discuss how to improve image classification algorithms given a large dataset as CIFAR-100 [7]. Image classification is still a very challenging task that lacks interpretability yet a much needed computer vision task requiring precision in this era of advanced Deep Learning algorithms. We discuss the methods that sheds light inside the algorithms and how they classify these images along with the reasons confusing the learning process of them. We used GRADCAM [10] to help this process and discussed the results accordingly.

1. Introduction

Since the breakthrough of Convolution Networks, image classification has become a very increasingly faster task to perform but also one of the most classic ML task lacks human understanding. The complexity of algorithms have increased dramatically in the past years [**?**] without properly shedding light on how these models are arriving at their predictions. This causes hindrance in improving model performance as we don't have a deeper understanding of the models. Furthermore, this hinders the process of industrially incorporating these models without the proper explanation of how they would impact real-world situations.

In recent times, there are many counterfactual algorithms trying to elucidate these models [12] amongst which Grad-CAM [10] is a visual tool that helps to understand these models by producing heatmaps of the predictions of deep layers and their extracted features. These can be used to understand not only the model's interpretability but also the proper use of neeeded image information. It helps us illustrate how the models arrive at their predictions and what leads the model to misclassification of images. Understanding this process is essential to modify the initial data collection process and tune the hyperparameters to improve performance of a particular model without needlessly introducing complex architectures. In this paper, we share our insights of various image classification algorithms using Grad-CAM on the Cifar-100 dataset.

2. Related work

Many attempts have been made to explain the black-box nature of Deep Learning algorithms and ethically establish their use in real-world scenerios [1] [12]. Visual explanatory methods are one of the popular methods that falls under the category of feature(pixel) importance explanation such as LFI-CAM [8], Score-CAM [13], LIFT-CAM [6] and others. These models require the weights of the model to visualize the inner layers and reflect on how they can be used to understand the flow of decision-making process of these models.

From industry to very sensitive data like medical data has been specially experimented to determine the interpretability of these predictions. One such is ECGrad-CAM that allows analysis of these algorithms trained on ECG features using attention maps [4]. In these cases, it is important to understand the situation the problem statement is being applied to. This is why, in our work we chose a standard image classification dataset which is also relatively large. There is another variant of gradient based method named Grad-CAM++ [**?**], due to time limitations we keep our work focused on Grad-CAM.

3. Method

As Grad-CAM requires re-training and we wanted to get insights on the CIFAR-100 dataset, we trained a number of models of scratch such as ResNet, EfficientNet, SwinTransformer, MobileNet, and RegNet. These well-recognized models are often used in theory but due to the lack of explainibilty, are not widely used in real-world. We have trained these models (Tab. 1) with the help of different optimizers, learning rate schedulers, and other hyperparameters. Finally, Grad-CAM has been used for the various outcomes from these models to identify the features that each model is looking for in the image.

3.1. Dataset

A subset of the Tiny Images dataset, the CIFAR-100 [7] dataset contains 60000 32x32 color images. The CIFAR-100's 100 classes are divided into 20 superclasses. Per class, there are 600 pictures. Each image has a "fine" and a "coarse" label, indicating the class (and the superclass) to which it belongs. For each class, there are 100 testing images and 500 training images.

3.2. Algorithms

EfficientNet [11] stands out from other deep neural architectures by using an uniform scaling of parameters that simplifies model tuning tremendously. The authors also state that scaling the architecture in this way lead to a stronger family of models that heighten the model performance without actively performing a grid search to build deeper models. EfficientnetV2 [**?**] originates from the same family with a small and faster architecture as proven by the authors. The inference time is reduced down to 24ms and training time as low as 7 hours which is significantly lesser (65%) than the original EfficientNet-B5.

MobileNetV3 originates from the Mobilenet family built for mobile phone CPUs with the help of Network Architecture Search and can perform image classification, segmentation and detection tasks [5]. This architectures is lightweight and claimed to have 3.2% better accuracy than its predecessor MobileNetV2 while being faster and less latent.

The ResNet family uses skip connections to pass information within deep layers known as residual learning. Architecture with as low as 34 layers can perform greatly on imagenet [3] and therefore, the pretrained models have gain much popularity in classificationt tasks. RegNet [14] improves upon the additive function of ResNet architecture and introduces a regulatory module to extract further features from the images. This simple-to-implement method is shown to perform well compared to classic ResNets.

Swin Transformers are vision transformers [2] that use transformers to solve image related tasks by creating patches of images. Swin transformers generate hierarchical feature maps of these patches using a shifted window scheme [9].

3.3. Grad-CAM

Grad-CAM‘ [10] is a technique for increasing the transparency of Convolutional Neural Network (CNN)-based models by showing the input areas that are "essential" for these models' predictions, also known as visual explanations. It creates a rough localization map that highlights the key areas in the picture for concept prediction by using the gradients of any target concept. It helps explaining the model's failure and identifying better models despite the same performance showed by conventional metrics.

Model	Params
EfficientnetV2	22M
Mobilenet	3.3M
RegNet	4.3M
ResNet34	21.3M
ResNet50	23.7M
ResNet101	42.7M
Swin Transformer	28.3M

Table 1. Discussed Deep Learning Algorithms

4. Experiment

We used 10% of the dataset as the validation set and 90% as the training set for this project. Since this is a classification problem, the loss function is cross-entropy loss. AdamW and stochastic gradient descent are the optimizers, and a batch size of 64 will be used for gradient descent. Gradient descent is approximated by stochastic gradient descent. It is significantly quicker to compute the gradient of the loss function when it is applied to a batch of all the training points rather than the entire set. The method is actually helped by the significant amount of noise that is introduced by this stochastic batch sampling of training data in order to keep it from becoming trapped in small local minima.

5. Results and Discussion

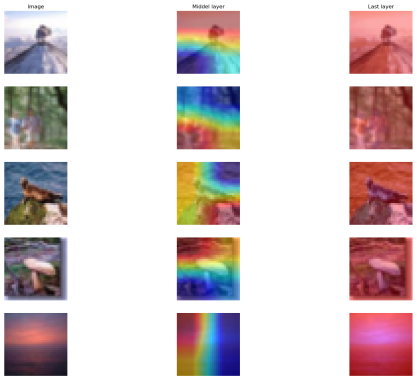

Figure 1. EfficientNet

In order to generate visual explanations, the grad cam has been applied to the intermediate layers and last layers.Grad-cam shows the areas that the network has considered important. In Figure 1 of the EfficietNet, Grad-cam cannot detect any features in the final layer, showing unreasonable predictions have plausible justifications since the im-

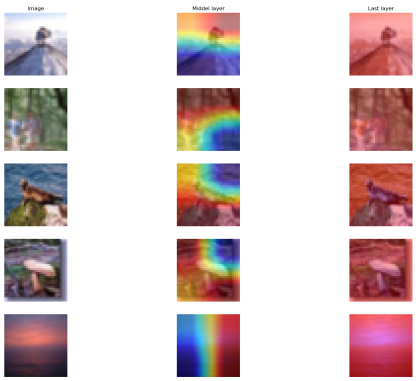

Figure 2. MobilenNet

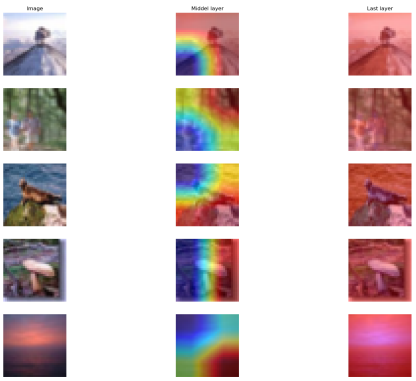

Figure 4. ResNet34

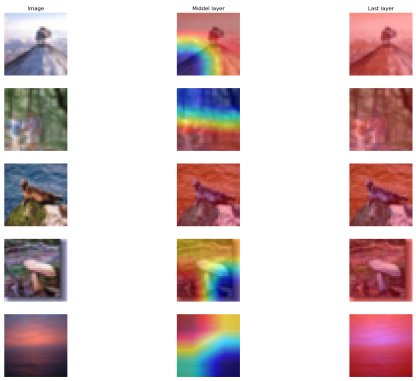

Figure 3. RegNet

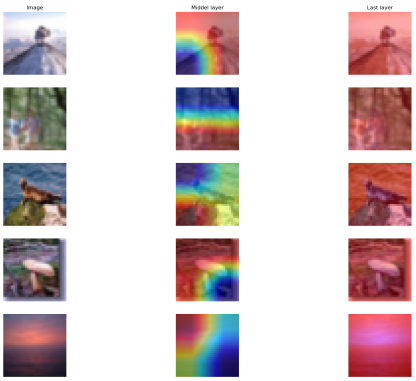

Figure 5. ResNet50

Model	Accuracy	Runtime
Efficientnet	83.65%	Frumpy
Mobilenet	62.85%	Frobbly
RegNet	75.47%	
ResNet34	70.41%	
ResNet50	87.15%	
ResNet101	68.48%	
Swin Transformer	70.85%	

Table 2. Accuracy for 40 epochs on Cifar-100 dataset

age size needed is not sufficient. Despite not having undergone much training, it is surprising that the model can recognize certain features in the middle layers. For ResNet, EfficientNet, RegNet, MobileNet, and Swin Transformer, we provide grad- cam visulaization.

6. Conclusion

In this research, we attempted to use Grad-CAM on CIFAR-100 using a number of different models. Each model has been carefully adjusted and trained for 40 iterations. Because the last layer's CIFAR-100 picture is very small, Grad-CAM results are not very significant. All of the Grad-Cam models' output and accuracy have been compared.

References

[1] Alejandro Barredo Arrieta, Natalia Diaz Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado González, Salvador García, Sergio Gil-López, Daniel Molina, V. Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 10 2019. 1

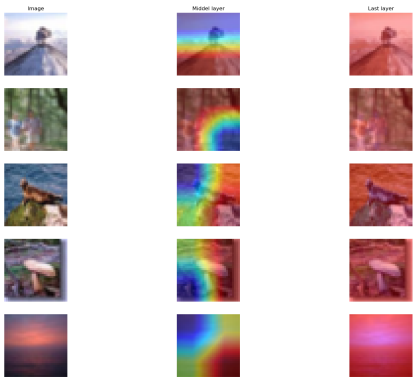

Figure 6. MobilenNet

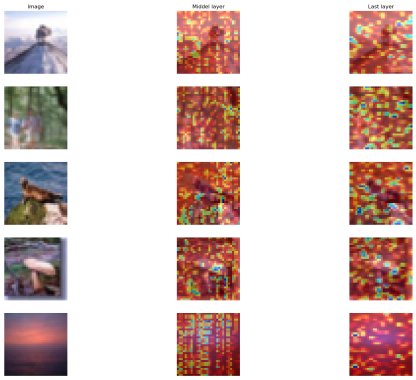

Figure 7. Swin Transformer

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 10 2020. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 06 2016. 2

[4] Steven Hicks, Jonas Isaksen, Vajira Thambawita, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Inga Strumke, Christina Ellervik, Morten Olesen, Torben Hansen, Claus Graff, Niels-Henrik Holstein-Rathlou, Pål Halvorsen, Mary Maleckar, Michael Riegler, and Jorgen Kanters. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Scientific Reports*, 05 2021. 1

[5] Andrew Howard, Ruoming Pang, Hartwig Adam, Quoc Le, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, and Yukun Zhu. Searching for mobilenetv3. pages 1314–1324, 10 2019. 2

[6] Hyungsik Jung and Youngrock Oh. Lift-cam: Towards better explanations for class activation mapping, 02 2021. 1

[7] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). 1, 2

[8] Kwang Lee, Chaewon Park, Junghyun Oh, and Nojun Kwak. Lfi-cam: Learning feature importance for better visual explanation, 05 2021. 1

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. pages 9992–10002, 10 2021. 2

[10] Ramprasaath Rs, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128, 02 2020. 1, 2

[11] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 05 2019. 2

[12] Tom Vermeire and David Martens. Explainable image classification with evidence counterfactual, 04 2020. 1

[13] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. pages 111–119, 06 2020. 1

[14] Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. Regnet: Self-regulated network for image classification, 01 2021. 2