Dilated Convolution with Learnable Spacings makes visual models more aligned with humans: a Grad-CAM study

Rabih Chamas[1] , **Ismail Khalfaoui-Hassani**[3] , **Timothée Masquelier**[2,3]

[1]LIS, CNRS.

[2]CerCo UMR 5549, CNRS. [3]Université de Toulouse, France.

rabih.chamas@lis-lab.fr, ismail.khalfaoui-hassani@univ-tlse3.fr, timothee.masquelier@cnrs.fr

Abstract

Dilated Convolution with Learnable Spacing (DCLS) is a recent advanced convolution method that allows enlarging the receptive fields (RF) without increasing the number of parameters, like the dilated convolution, yet without imposing a regular grid. DCLS has been shown to outperform the standard and dilated convolutions on several computer vision benchmarks. Here, we show that, in addition, DCLS increases the models' interpretability, defined as the alignment with human visual strategies. To quantify it, we use the Spearman correlation between the models' Grad-CAM heatmaps and the ClickMe dataset heatmaps, which reflect human visual attention. We took eight reference models – ResNet50, ConvNeXt (T, S and B), CAFormer, ConvFormer, and FastViT (sa_24 and 36) – and drop-in replaced the standard convolution layers with DCLS ones. This improved the interpretability score in seven of them. Moreover, we observed that Grad-CAM generated random heatmaps for two models in our study: CAFormer and ConvFormer models, leading to low interpretability scores. We addressed this issue by introducing Threshold-Grad-CAM, a modification built on top of Grad-CAM that enhanced interpretability across nearly all models. The code and checkpoints to reproduce this study are available at: https://github.com/rabihchamas/DCLS-GradCAM-Eval

1 Introduction

Deep learning neural networks are extremely powerful for a myriad of tasks, including image classification. However, despite being very powerful tools, they remain black box models, and understanding how they arrive at their results can be a major challenge. Explainability methods in deep learning aim to explain why a particular model predicts a particular result.

One application where explainability methods have been successfully in use for several years is image classification. The most popular and successful models nowadays for image classification include convolution and/or attention layers.

When a model contains only convolutions, it is called a fully convolutional neural network or CNN, when it contains only multi-head self-attention (MHSA) layers, it is called a transformer or, in the context of computer vision, a vision transformer, and when a model contains both layers, it is called a hybrid model.

Despite their very high accuracy, most of these models remain very opaque, and the lack of explicability of the latter, especially in computer vision, raises concerns about trust, fairness, and interoperability, hindering their adoption in sensitive areas such as medical diagnosis [Collenne et al., 2024] or autonomous vehicles.

This also applies to recent advances such as Dilated Convolution with Learnable Spacings (DCLS) [Khalfaoui-Hassani et al., 2023b], which shows promising performance gains in tasks such as image classification, segmentation, and object detection.

While the accuracy of DCLS is encouraging, its black-box nature demands attention. Thus, we have been motivated to explore explainability measures and scores specifically for DCLS, with the hope of shedding light on its underlying decision-making processes.

The taxonomies used for explainability in the artificial intelligence field of research are diverse and constantly evolving as new approaches are discovered. However, a common way to proceed is to distinguish between two major families of model explainability methods: global methods and local methods [Speith, 2022; Schwalbe and Finzel, 2023].

Global methods describe the overall behavior of the model, considering general patterns and the importance of features. Some examples include Partial Dependence Plots (PDPs) [Friedman, 2001] and SHapley Additive explanations (SHAP) [Lundberg and Lee, 2017]. Local methods, on the other hand, focus on explaining individual predictions, focusing on why the model made a particular decision for a particular input. Examples include Local Interpretable Model-Agnostic Explanations (LIME) [Ribeiro et al., 2016] and Gradient-weighted Class Activation Mapping (Grad-CAM) [Selvaraju and et al., 2017].

Grad-CAM is a popular method that helps visualize which parts of an image are most important to the model's decision. For the needs of our study, we designed a new explainability method based on Grad-CAM that we called Thershold-Grad-CAM. This new explainability method overcomes some is-

sues tied to the failure of traditional Grad-CAM, in particular, for ConvFormer and CAFormer architectures [Yu *et al.*, 2023].

The objective of this paper is, on the one hand, to perform a comparative study in terms of explainability scores between recent state-of-the-art models in computer vision, namely ConvNeXt [Liu *et al.*, 2022], ConvFormer [Yu *et al.*, 2023], CAFormer [Yu *et al.*, 2023], and FastViT [Vasu *et al.*, 2023] in their original form, and, on the other hand, to perform the same study between these same models and their DCLS-enhanced counterparts.

What motivated the comparative study presented here in this paper is the qualitative similarity we noticed between human attention heatmaps obtained from the ClickMe dataset [Linsley *et al.*, 2019] and those obtained by models empowered with DCLS. Figure 1 gives an overview of this similarity based on the ConvNeXt-B model. The images presented in the figure 1 were selected from the ClickMe dataset. To help illustrate our point, we have selected a few images where the heatmaps are visually relevant. We then quantitatively confirmed this remarkable alignment of DCLS models and human attention heatmaps through a rigorous study of the Spearman correlation [Zar, 2005] between heatmaps generated by Threshold-Grad-CAM and heatmaps made by human participants in the ClickMe dataset.

We will refer to a model empowered with DCLS by its original name followed by the suffix: "_dcls". We create a DCLS-empowered model by performing a drop-in replacement of all the depthwise separable convolutions of this model [Chollet, 2017; Sandler *et al.*, 2018] with DCLS ones.

DCLS was introduced in Khalfaoui-Hassani *et al.* [2023b], where it exhibited better performance than the depthwise separable convolution and the dilated convolution [Yu and Koltun, 2015] for computer vision tasks such as image classification, semantic segmentation, and object detection, as well as for computer audition tasks such as audio classification [Khalfaoui-Hassani *et al.*, 2023a]. Initially, the DCLS method used bilinear interpolation, in Khalfaoui-Hassani *et al.* [2023c] this interpolation was extended to Gaussian. DCLS has focused on learning the positions of kernel elements along their weights. We believe this advance is important for tasks that require a nuanced understanding of visual context, similar to human perception.

2 Methods

2.1 ClickMe dataset

To quantitatively evaluate the interpretability of the models, we employed the ClickMe dataset [Linsley *et al.*, 2019], which was introduced to capture human attention strategies in classification tasks. The dataset collection process involved a single-player online game, ClickMe.ai [Linsley *et al.*, 2019], where players identified the most informative parts of an image for object recognition. The alignment of model-generated heatmaps with those from the ClickMe dataset measures how closely a model's attention strategy mirrors human strategy.

2.2 DCLS method

Although larger convolution kernels can improve performance, increasing the kernel size increases the number of parameters and computational cost. Yu and Koltun [2015] introduced dilated convolution (DC) to expand the kernel without increasing parameters. DC inserts zeros between kernel elements, effectively enlarging the kernel without adding new weights. However, DC uses a fixed grid, which can limit performance.

Khalfaoui-Hassani *et al.* [2023b] presented DCLS as a new method that builds upon DC. Instead of using fixed spacings between non-zero elements in the kernel, DCLS allows learning these spacings through backpropagation. An interpolation technique is used to overcome the discrete nature of the spacings while maintaining the differentiability necessary for backpropagation.

2.3 Grad-CAM and Threshold-Grad-CAM

Grad-CAM is a technique that provides visual explanations for the decisions made by deep neural networks. The method uses the gradients of a target concept, propagated into the final convolutional layer of a deep neural network, to produce a localization map highlighting the regions in the input image that are crucial for predicting this concept [Selvaraju and et al., 2017]. Grad-CAM adapts to various network architectures by focusing on the last layer of interest before a classification head or pooling operation. The method is detailed in the supplementary material.

Threshold-Grad-CAM

In the standard implementation of Grad-CAM, a ReLU activation is applied to the weighted combination of activation maps post-summation. This is predicated on the assumption that positive features should be exclusively highlighted as they are the ones contributing to the class prediction [Selvaraju and et al., 2017]. However, our observations suggest that applying ReLU after the summation can inadvertently suppress useful signals when negative activations are present, as they may negate some positive activations when summed. This phenomenon becomes particularly pronounced in architectures such as ConvFormer and CAFormer, where we observed that the resulting heatmaps were no more informative than random heatmaps. We believe this is due to the choice of a specific activation: StarReLU in these two architectures Yu *et al.* [2023], which depends on two learnable parameters: scale and bias.

To address this issue, we propose applying ReLU to the activation maps before their summation. We then normalize the heatmaps. Finally, the heatmaps are thresholded to retain values above a predetermined threshold (determined experimentally to be 0.3 for optimal results on the ClickMe dataset). The modified Grad-CAM process is described in the supplementary material. Our experiments demonstrated that this modification significantly improved the interpretability of the heatmaps generated for ConvFormer and CAFormer.

3 Related work

The field of interpretable and explainable AI has recently gained significant attention within the AI community. Exten-
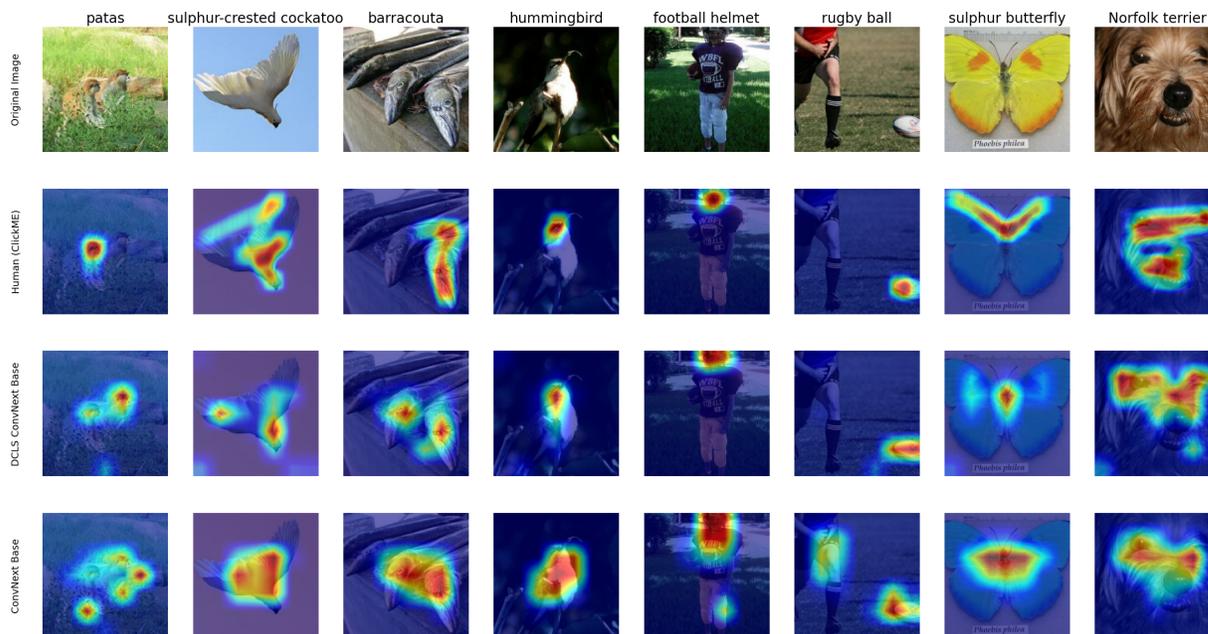
Figure 1: Visualization of Heatmaps on ClickMe dataset Images. First row: original images from the ClickMe dataset. Second row: the same images superimposed with heatmaps created by humans from the ClickMe project. Third row: Threshold-GradCAM heatmaps of the ConvNeXt base model enhanced with DCLS. Fourth row: Threshold-GradCAM heatmaps of the baseline ConvNeXt base model without DCLS.

sive research efforts range from defining key terms such as interpretability and explainability to developing explainability methods assessing their trustworthiness and evaluating the interpretability of deep learning models. Gilpin *et al.* [2018] distinguished between interpretability and explainability and highlighted the challenge of achieving complete and interpretable explanations at the same time. Doshi-Velez and Kim [2017] defined interpretability as the ability to present model decisions in terms understandable to humans. In their study, Mohseni *et al.* [2021] utilized multi-layer human attention masks to benchmark the effectiveness of explanation methods such as Grad-CAM and LIME. Velmurugan *et al.* [2020] proposed functionally grounded evaluation metrics that assess the trustworthiness of explainability methods, including LIME and SHAP. Furthermore, Fel *et al.* [2022] employed the ClickMe dataset to investigate the alignment between human and deep neural network (DNN) visual strategies, applying a training routine that aligns these strategies, as a result enhancing categorization accuracy. In our study, we align with the interpretability definitions in the literature. We employ human heatmaps from the ClickMe dataset as ground truth to evaluate our model's interpretability.

4 Experiments

In this section, we present the experimental setup used to compare the performance and interpretability of our proposed models. Specifically, we calculated the top-1 accuracy of the models trained on the ImageNet1k validation dataset [Deng and et al., 2009] to assess their classification effectiveness. For interpretability, we employed Spearman's correlation as

a metric to compare the alignment between human-generated heatmaps from the ClickMe dataset and the model-generated heatmaps. We assessed the interpretability of the heatmaps produced using two different methods: Grad-CAM and our proposed Threshold-Grad-CAM.

4.1 Results

We present the results of integrating DCLS into state-of-the-art neural network architectures and our novel update to the Grad-CAM technique. Our experiments evaluated model interpretability, which we defined as the degree of alignment between heatmaps generated by explainability methods and those derived from human visualization strategies.

4.2 Improvement in Model Interpretability with DCLS

Our experiments incorporated DCLS into five model architectures: ResNet, ConvNeXt, CAFormer, ConvFormer, and FastViT. We trained each model on ImageNet1k. When this is mentioned by _dcls, it means that the training has been done by replacing each depth-separable convolution of the baseline model with DCLS.

The results showed an enhancement in model interpretability with all models but FastViT_sa24. When equipped with DCLS, ConvNeXt improved in heatmap alignment. The score improved with both Grad-CAM and Threshold-Grad-CAM methods, as shown in Table 1 and in Figure 2.

Since Grad-CAM generates random heatmaps on CAFormer and ConvFormer architectures, we experimented with Threshold-Grad-CAM. Similar to ConvNeXt, CaFormer and ConvFormer showed higher interpretability scores when

Table 1: Interpretability scores of various models on the ClickMe dataset using GradCAM and Threshold-GradCAM, with and without DCLS. The table presents the top-1 accuracy of each model alongside their respective interpretability scores. Models with the "_dcls" suffix indicate the use of DCLS.

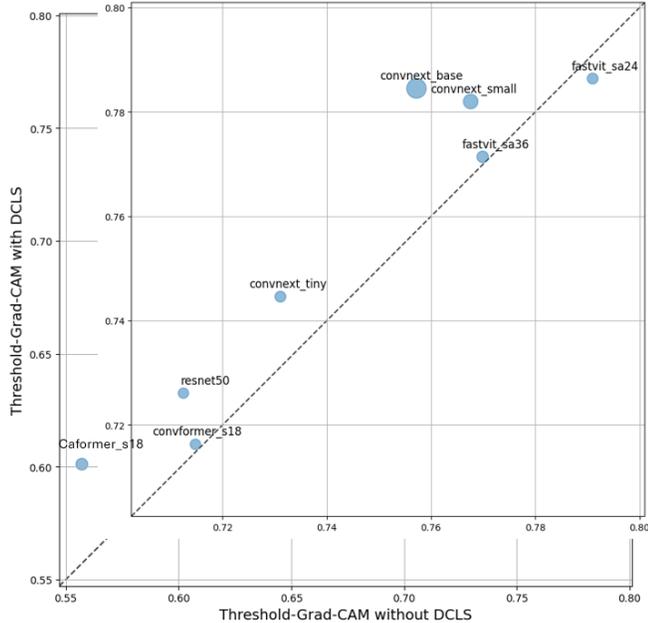| Model	Top1-accuracy	Grad-CAM score	Threshold-Grad-CAM score
convnext_tiny	82.1	0.6696	0.7311
convnext_tiny_dcls	82.48	0.7561	0.7446
convnext_small	83.15	0.7417	0.7676
convnext_small_dcls	83.72	0.788	0.782
convnext_base	83.83	0.7565	0.7572
convnext_base_dcls	84.09	0.7979	0.7845
fastvit_sa24	81.11	0.7608	0.7933
fastvit_sa24_dcls	82.48	0.7511	0.7754
fastvit_sa36	82.85	0.6699	0.7699
fastvit_sa36_dcls	82.69	0.699	0.7714
caformer_s18	83.66	0.1719	0.5571
caformer_s18_dcls	83.56	-0.0594	0.6011
convformer_s18	82.98	-0.0375	0.7148
convformer_s18_dcls	83.06	-0.1285	0.7163
resnet50	77.84	0.6135	0.7125
resnet50_dcls	78.35	0.6252	0.7261

Figure 2: Comparison of models interpretability score using Threshold-GradCAM with and without DCLS. Each point represents a different model, plotted according to its interpretability score without DCLS on the x-axis and with DCLS on the y-axis. Models above the dashed line demonstrate improved performance with the inclusion of DCLS.

used with DCLS. The FastViT_sa24 model showed a high interpretability score, even without incorporating DCLS, and applying DCLS didn't improve the score.

In addition, DCLS increased the top-1 accuracy in all models but CAFormer_s18 and FastViT_sa36 (Table 1).

5 Discussion

Except for FastViT, all model families studied show two points: first, an increase in accuracy when the depthwise separable convolution is replaced by DCLS, and second, an increase in the Treshold-Grad-CAM explanability score when this same modification is made. FastViT is a special model because the test inference is performed with a kernel reparametrization identical to that of RepLKNet Ding *et al.* [2022]. This could interfere with the DCLS method, which is in fact a different reparametrization, and might explain why the results for this family of models were not correlated in the same way as for the other studied models. Furthermore, the results presented here are significant since we tested three different training seeds for the ConvNeXt-T-dcls model and found an accuracy of 82.49 ± 0.04 and a Treshold-Grad-CAM score of 0.7466 ± 0.004.

Fel *et al.* [2022] utilized the ClickMe dataset to compare human and DNN visual strategies on ImageNet [Deng and et al., 2009]. They adopted a classic explainability method: Image Feature Saliency [Jiang *et al.*, 2015], to generate comparable feature importance maps for 84 deep neural networks (DNNs). They report that as DNNs become more accurate, a trade-off emerges where their alignment with human visual strategies starts to decrease. In contrast, our study employs Grad-CAM for analyzing DNN visual strategies. Unlike the findings of Fel *et al.* [2022], our use of Grad-CAM did not reveal such a trade-off. We think that this discrepancy is due to the differences in the explanatory methods used, which highlights the influence of analytical tools in interpreting DNN visual strategies.

Furthermore, it is conceivable that models with higher interpretability scores may focus more on those image features mostly correlated with the label class. A preliminary examination of the ClickMe dataset reveals that humans tend to concentrate solely on the object representing the class label within the image, ignoring other less directly related features to the class. This behavior likely stems from a nuanced human understanding of the concepts. Therefore, alignment with human-generated heatmaps might reflect a model's robustness.

6 Conclusion

In this study, we investigated the interpretability of recent deep neural networks using Grad-CAM-based methods for image classification tasks. We found that employing Dilated Convolution with Learnable Spacings enhances network interpretability. Our results indicate that DCLS-equipped models better align with human visual perception, suggesting that such models effectively capture conceptually relevant features akin to human understanding. Future work could focus on investigating the explainability score of DCLS using black-box methods such as RISE.

References

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017.

Jules Collenne, Jilliana Monnier, Rabah Iguernaissi, Motasem Nawaf, Marie-Aleth Richard, Jean-Jacques Grob, Caroline Gaudy-Marqueste, Séverine Dubuisson, and Djamal Merad. Fusion between an algorithm based on the characterization of melanocytic lesions' asymmetry with an ensemble of convolutional neural networks for melanoma detection. *Journal of Investigative Dermatology*, 2024.

Jia Deng and et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *stat*, 1050:2, 2017.

Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, 35:9432–9446, 2022.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

Ismail Khalfaoui-Hassani, Timothée Masquelier, and Thomas Pellegrini. Audio classification with dilated convolution with learnable spacings. In *NeurIPS 2023 Workshop on Machine Learning for Audio*, 2023.

Ismail Khalfaoui-Hassani, Thomas Pellegrini, and Timothée Masquelier. Dilated convolution with learnable spacings. In *The Eleventh International Conference on Learning Representations*, 2023.

Ismail Khalfaoui-Hassani, Thomas Pellegrini, and Timothée Masquelier. Dilated convolution with learnable spacings: beyond bilinear interpolation. In *ICML 2023 Workshop on Differentiable Almost Everything: Differentiable Relaxations, Algorithms, Operators, and Simulators*, 2023.

Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. In *International Conference on Learning Representations*, 2019.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Sina Mohseni, Jeremy E Block, and Eric Ragan. Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, page 22–31, New York, NY, USA, 2021. Association for Computing Machinery.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023.

Ramprasaath R Selvaraju and et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Timo Speith. A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, 2022.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5785–5795, 2023.

Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Evaluating explainable methods for predictive process analytics: A functionally-grounded approach, 2020.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Jerrold H Zar. Spearman rank correlation. *Encyclopedia of Biostatistics*, 7, 2005.

A   Appendix: Grad-CAM Implementation

Algorithm 1 Grad-CAM Implementation

Input: Image I, Target class c, Trained Convolutional Neural Network CNN
Output: Heatmap H visually highlighting influential regions for class c

1: **Forward Pass:**
   Process image I through CNN to obtain feature maps at the last convolutional layer A. Let A^k be the feature map for the k-th channel.
2: **Compute Gradients:**
   Compute the gradient of the loss for class c, denoted y^c, with respect to the feature maps A, resulting in $\frac{\partial y^c}{\partial A^k}$.
3: **Global Average Pooling of Gradients:**
   For each feature map channel k, compute the global average of the gradients:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

   where i, j index spatial dimensions and Z is the number of elements in A^k.
4: **Weighted Combination of Feature Maps:**
   Compute the weighted sum of the feature maps using the weights α_k^c:

$$L^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right)$$

5: **Generate Heatmap:**
   Resize L^c to the size of the input image I to get the heatmap H.
6: **Overlay Heatmap on Original Image:**
   Superimpose H onto the original image I for visualization, adjusting the transparency to ensure visibility of underlying features.

B   Appendix: Threshold-Grad-CAM Implementation

Algorithm 2 Threshold GradCAM

Input: Weighted activation maps A^k
Parameter: Threshold value $t = 0.3$
Output: Final heatmap H

1: **Apply ReLU Activation:**
   Apply the ReLU function to each weighted activation map to filter out negative values. This step prevents the cancellation of positive activations during summation:

$$A_{\text{ReLU}}^k = \text{ReLU}\left(\alpha_k^c A^k \right)$$

2: **Summation of Activated Maps:**
   Sum the ReLU-activated maps: $S = \sum_k A_{\text{ReLU}}^k$
3: **Normalization:**
   Normalize the summed activation map S to ensure values are scaled consistently:

$$N = \frac{S}{\max(S)}$$

4: **Apply Thresholding:**
   Apply a threshold of t to reduce noise and enhance the focus on relevant regions:

$$H = \begin{cases} N & \text{if } N \geq t \\ 0 & \text{otherwise} \end{cases}$$

Our revised approach yields more coherent and focused visual explanations, as validated by quantitative assessments of the ClickMe dataset.

C Appendix: DCLS vs. Baseline: Interpretability Analysis with Grad-CAM and Threshold-Grad-CAM
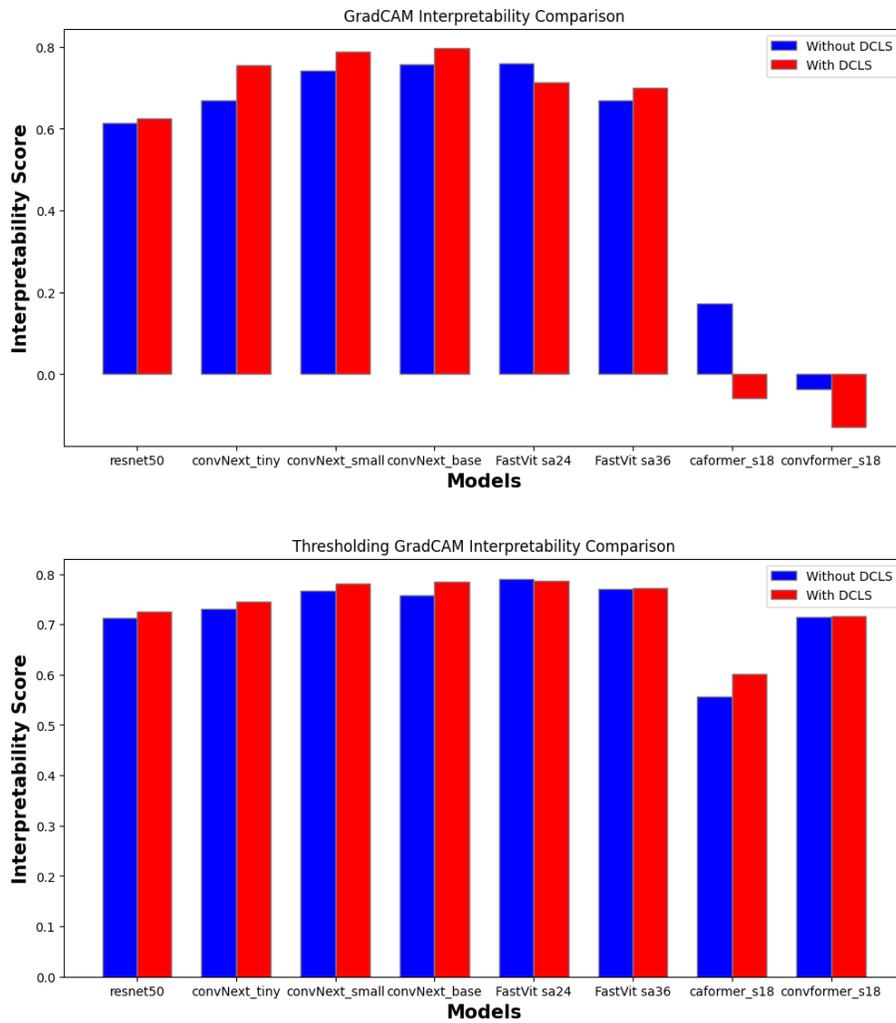

Figure 3: Comparative analysis of interpretability scores across different models using Grad-CAM and Threshold-Grad-CAM techniques. Top: The interpretability scores with Grad-CAM. Bottom: The interpretability scores with Threshold-Grad-CAM. Both subfigures highlight the difference in scores with and without DCLS. The results indicate that DCLS generally improves interpretability scores for most models.

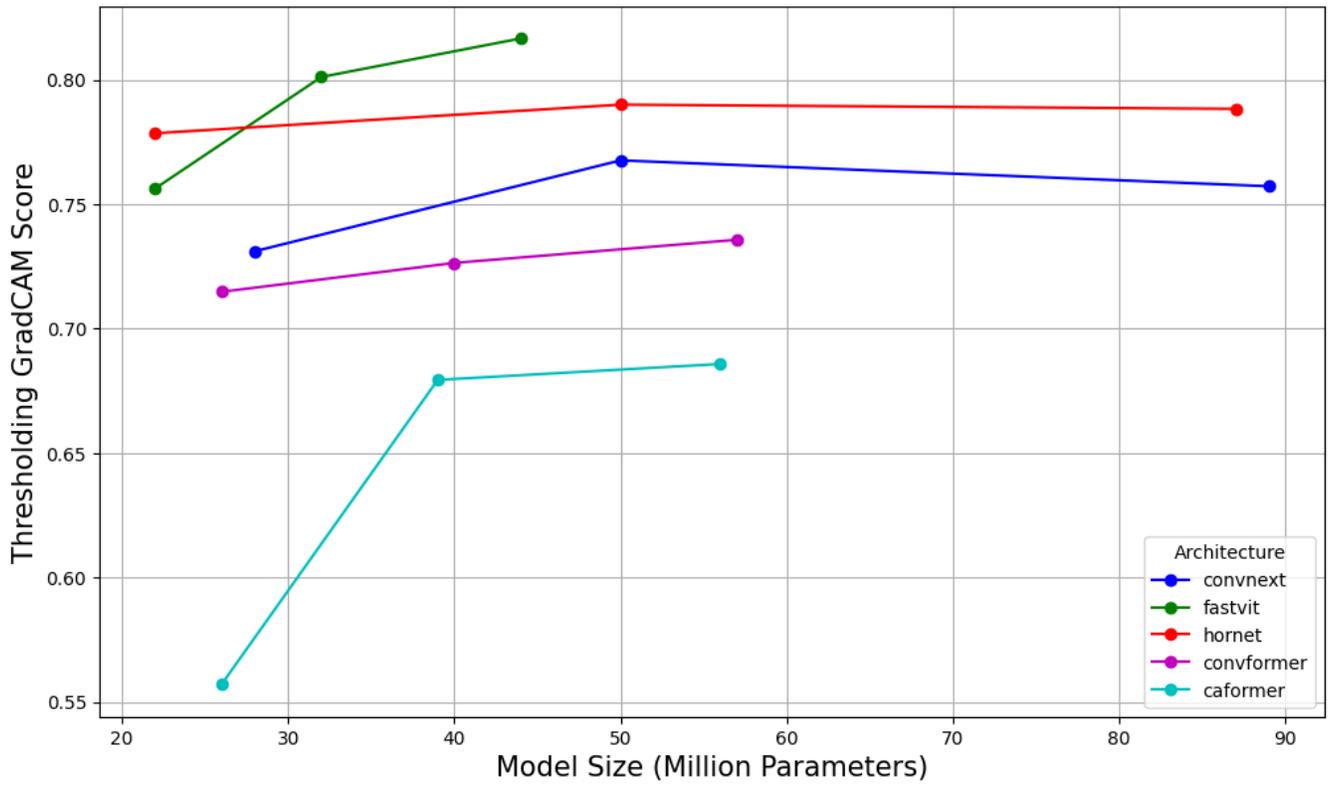## D   Appendix: Model Size vs. Interpretability Score Using Threshold-Grad-CAM

Figure 4: Correlation between model size and Interpretability for baseline models, using Threshold-Grad-CAM scores. Larger models tend to have higher interpretability scores, suggesting a positive correlation between model size and explainability in baseline models.

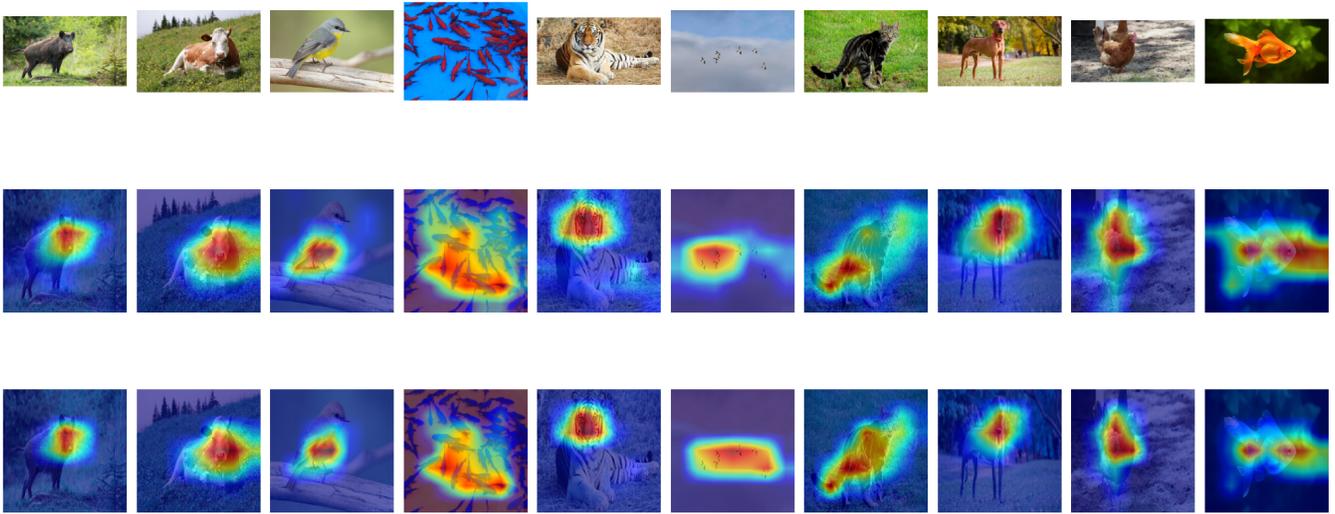# E   Appendix: Visualizing Grad-CAM and Threshold-Grad-CAM Heatmaps

Figure 5: ResNet50 Grad-CAM heatmaps and Threshold-Grad-CAM heatmaps across 10 randomly chosen license-free internet images. Top row: Original images. Middle row: Images with Grad-CAM heatmaps. Bottom row: Images with Threshold-Grad-CAM heatmaps.
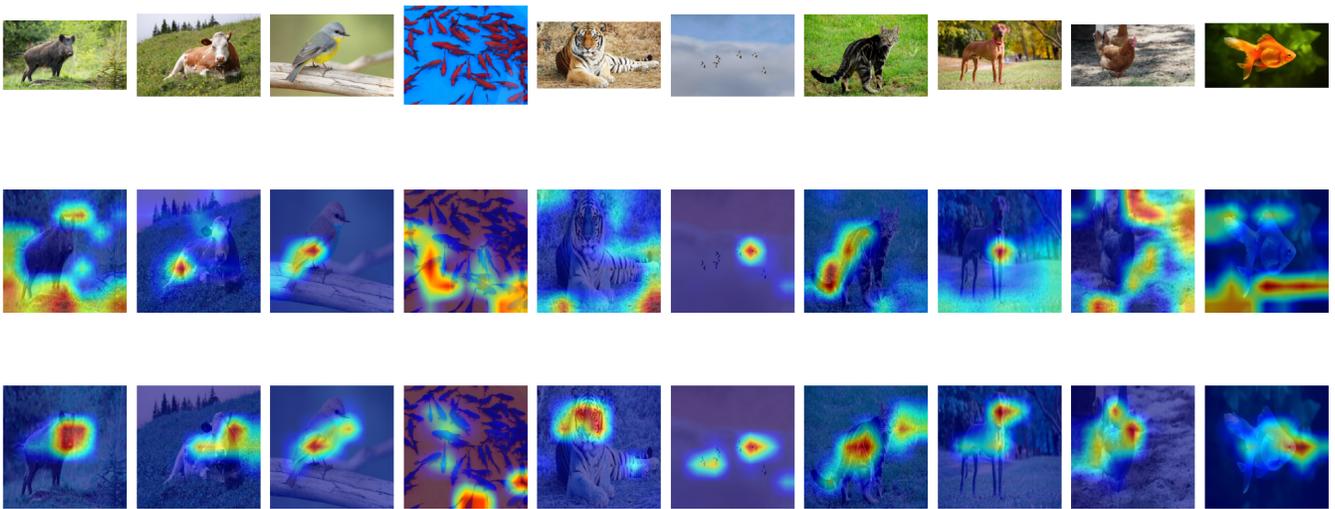

Figure 6: ConvFormer Grad-CAM heatmaps and Threshold-Grad-CAM heatmaps across 10 randomly chosen license-free internet images. Top row: Original images. Middle row: Images with Grad-CAM heatmaps. Bottom row: Images with Threshold-Grad-CAM heatmaps.