KSÜ Mühendislik Bilimleri Dergisi, 27(3), 2024
Araştırma Makalesi

KSU J Eng Sci, 27(3), 2024
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 02.02.2024
Kabul Tarihi : 22.07.2024

Received Date : 02.02.2024
Accepted Date : 22.07.2024

HAYCAM VS EIGENCAM FOR WEAKLY-SUPERVISED OBJECT DETECTION ACROSS VARYING SCALES

FARKLI ÖLÇEKLERDE ZAYIF DENETİMLİ NESNE TESPİTİ İÇİN HAYCAM VE EIGEN KARŞILAŞTIRILMASI

Ahmet Haydar ORNEK[1] (ORCID: 0000-0001-7254-9316)
Murat CEYLAN[2*] (ORCID: 0000-0001-6503-9668)

[1] Huawei Türkiye R&D Center, Service Application DC, Istanbul 34764, Türkiye
[2] Konya Technical University, Electrical & Electronics Engineering Department, Konya 42130, Türkiye

* Sorumlu Yazar / Corresponding Author: Ahmet Haydar ORNEK, ahmet.haydar.ornek2@huawei.com

ABSTRACT

When a classification process is performed using Class Activation Maps, which is one of the Explainable Artificial Intelligence approaches, the areas influencing the classification on the input image can be revealed. In other words, it is demonstrated which part of the image the classifier model looks at to make a decision. In this study, a 200-class classification model was trained using the open-source dataset CUB 200 2011, and the classification results were visualized using the EigenCAM and HayCAM methods. When comparing object detection performances based on the areas influencing classification, the EigenCAM method reaches an IoU (Intersection over Union) value of 30.88%, while the HayCAM method reaches a value of 41.95%. The obtained results indicate that outputs derived using Principal Component Analysis (HayCAM) are better than those obtained using Singular Value Decomposition (EigenCAM).

Keywords: Explainable artificial intelligence, activation map, deep learning, eigencam, haycam

ÖZET

Açıklanabilir Yapay Zeka yaklaşımlarından biri olan Sınıf Aktivasyon haritaları ile bir sınıflama işlemi gerçekleştirildiği zaman, giriş görüntüsü üzerindeki sınıflamaya etki eden alanlar ortaya çıkarılabilmektedir. Yani bir sınıflayıcı modelin görüntünün hangi kısmına bakarak karar verdiği gösterilmektedir. Bu çalışmada açık kaynak bir veri seti olan CUB 200 2011 kullanılarak 200 sınıflı bir sınıflama modeli eğitilmiş ve sınıflama sonuçları EigenCAM ve HayCAM yöntemleri kullanılarak görselleştirilmiştir. Sınıflamaya etki eden alanlar kullanılarak gerçekleştirilen nesne tanıma performansları karşılaştırıldığında EigenCAM yöntemi %30.88 IoU değerine ulaşırken HayCAM yöntemi %41.95 değerine ulaşmaktadır. Elde edilen çıktılar Temel Bileşenler Analizi kullanılarak elde edilen sonuçların (HayCAM), Tekil Değer Ayrışımı kullanılarak elde edilen sonuçlardan (EigenCAM) daha iyi olduğunu göstermektedir.

Anahtar Kelimeler: Açıklanabilir yapay zeka, aktivasyon haritası, derin öğrenme, eigencam, haycam

ToCite: ORNEK, A., & CEYLAN, M., (2024). HayCAM vs EigenCAM for Weakly-Supervised Object Detection Across Varying Scales. *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, 27(3), 1077-1088.

INTRODUCTION

The rapid evolution of Artificial Intelligence (AI) has emphasized the importance of developing models that not only deliver accurate results but also provide insights into their decision-making processes (Krizhevsky et al., 2012). This has led to a growing focus on Explainable Artificial Intelligence (XAI) approaches, aiming to enhance transparency and interpretability (Arrieta et al., 2020). In this study, we explore the application of Class Activation Maps (CAMs) during the classification process as one such XAI approach. CAMs enable the identification of specific regions on the input image that significantly influence the classification decision, offering a better understanding of the model's reasoning (van der Velden et al., 2022).

With the XAI, CAMs serve as a crucial tool for visualizing areas in the input image that contribute most to the classification outcome. Specifically, we investigate the EigenCAM (Muhammad & Yeasin, 2020) and HayCAM (A. Ornek & Ceylan, 2022) methods, employing them in training a 200-class classification model on the CUB 200 2011 open-source dataset (Wah et al., 2011). Through extensive experimentation, our objective is to compare the performances of these visualization methods in revealing influential areas for accurate object detection.

The core of our exploration lies in the comparative analysis of object detection results obtained through EigenCAM and HayCAM. Utilizing the Intersection over Union (IoU) metric (D. Zhou et al., 2019) as a key indicator of alignment between predicted and ground truth bounding boxes, we observe a significant disparity. EigenCAM achieves a 30.88% IoU value, whereas HayCAM surpasses it with a notably higher value of 41.95% IoU. As shown in Figure 1, we leverage analytical techniques like Principal Component Analysis (PCA) (Wu et al., 2018) and Singular Value Decomposition (SVD) (Stewart, 1993) to underscore the superiority of insights derived from HayCAM over EigenCAM, particularly in the context of XAI. Both methods decrease the filter size from 512 to 1, but using different techniques. Although HayCAM originally reduced the number of filters to 19, in this study it was arranged to reduce the number of filters to 1 filter.

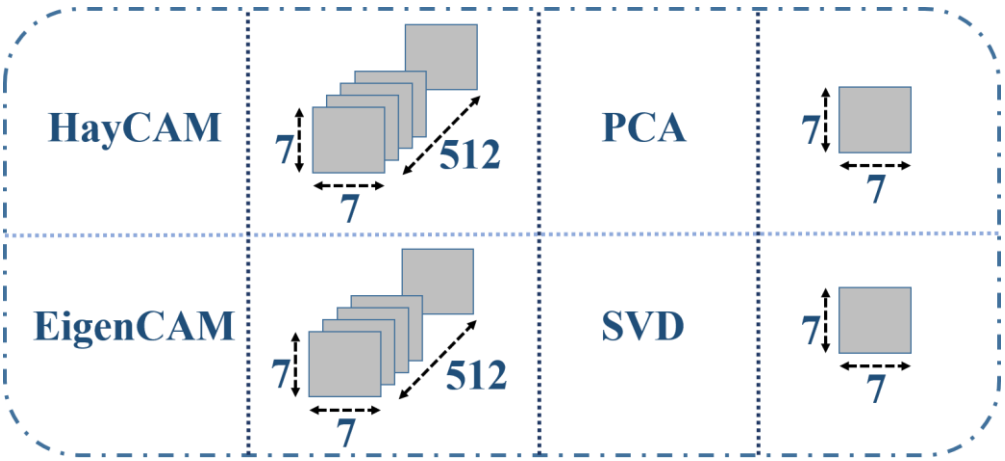

Figure 1. HayCAM and EigenCAM. While HayCAM Uses PCA to Decrease the Filters to 1 Filter, EigenCAM Uses SVD

The contributions of this study are threefold: First, we showcase the practical application of explainable AI in real-world datasets by utilizing the CUB 200 2011 open-source dataset in training a 200-class classification model. Second, the application of EigenCAM and HayCAM methods for visualizing classification results provides valuable insights into the decision-making process of the model, emphasizing crucial areas for accurate classification. Third, the comparison of object detection performance, based on areas influencing classification, reveals that HayCAM achieves a higher IoU.

The paper is structured as follows: Section 2 gives related works in XAI and Class Activation Maps. Section 3 details the dataset, and Section 4 describes the methodology, including the steps taken to train the classification model and visualize the results. Section 5 presents the experimental results, and Section 6 discusses the implications. Finally, Section 7 concludes the paper and outlines potential avenues for future research.

RELATED WORK

Visual XAI is a field focused on enhancing the transparency and interpretability of deep learning models (Ornek & Ceylan, 2022). It aims to unveil the decision-making processes of these models, fostering a better understanding of their inner workings (van der Velden et al., 2022). Various XAI methods, such as Class Activation Mapping (CAM) (B. Zhou et al., 2016), Gradient-weighted Class Activation Mapping (GradCAM) (Selvaraju et al., 2017), GradCAM++ (Chattopadhay et al., 2018), HayCAM (A. Ornek & Ceylan, 2022), and EigenCAM (Muhammad & Yeasin, 2020), generate activation maps highlighting image regions contributing to model decisions, thus facilitating comprehension of object identification mechanisms. These methods play a crucial role in promoting transparency and interpretability, building trust in responsible and ethical deep learning model usage.

GradCAM, a widely adopted visualization technique for understanding deep neural networks, generates heatmaps by computing gradients of CNN feature maps' output scores. Building on GradCAM, techniques like GradCAM++ consider second-order gradients for more precise heatmaps, capturing finer details. EigenCAM employs SVD to capture intrinsic feature map structures, producing robust heatmaps even amidst noise or perturbations.

CodCAM is an ensemble visual XAI method (Ornek & Ceylan, 2023), it uses the activation maps of different XAI methods such as GradCAM, and outputs a focused activation map. Although CodCAM provides a better activation mapping, its computational cost is higher than the single XAI methods.

HayCAM uses Principal Component Analysis (PCA) to further streamline filter sizes in the last convolutional layer. Empirically selecting filter sizes based on detected objects enables focused area extraction. Given that numerous filters may lead to scattered areas on images. Although HayCAM originally reduces the number of filters to 19, in this study, HayCAM reduces filters to one, enhancing focus on targeted areas during small object detection. CodCAM and HayCAM are two techniques that were developed in the doctoral dissertation of the first author (Ornek, 2023).

DATASET

CUB-200-2011 is an expanded iteration of CUB-200, a dataset comprising 200 distinct bird species (Wah et al., 2011). This dataset is widely regarded as a challenging benchmark for image classification tasks. The extended version of CUB-200-2011 amplifies the number of images per category by approximately twofold, thereby providing a more comprehensive and diverse collection of bird images. The extended version includes new part localization annotations, which enriches the dataset with more detailed and accurate information about the birds' anatomical features. Each image in the dataset is annotated with bounding boxes, part locations, and attribute labels, which makes the dataset a valuable resource for training and testing machine learning algorithms. Randomly selected images are shown in Figure 2.

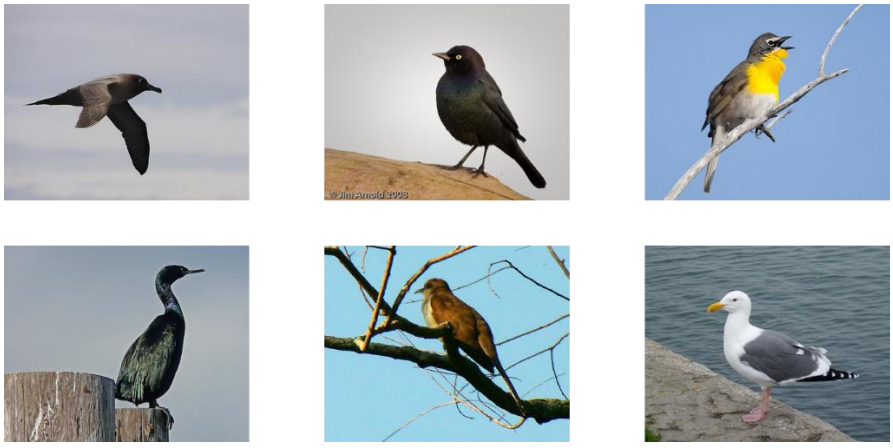

Figure 2. Sample Images from the Dataset

The foundation of our study rests on the utilization of the CUB 200 2011 dataset, a widely recognized and openly accessible dataset within the computer vision community. CUB 200 2011 comprises a diverse collection of images featuring 200 different bird species, each labeled with fine-grained details. 11780 images are used in this study. The counts of objects in 10 ratio intervals are presented in Table 1.

Table 1. Number Of Objects

Object/Image Ratio	Object Count
100% - 90%	522
90% - 80%	1290
80% - 70%	2173
70% - 60%	2767
60% - 50%	2558
50% - 40%	1752
40% - 30%	587
30% - 20%	126
20% - 10%	5
10% - 0%	0

Table 1 shows the distribution of images in the dataset based on their ratio. Between 100% and 90%, there are 522 images, between 90% and 80%, there are 1290 images, and so on. This dataset not only provides a rich variety of avian species but also presents a challenge for classification models due to the intricacies of distinguishing between closely related bird types. The inclusion of such a comprehensive dataset serves as a robust testing ground for evaluating the effectiveness of our XAI approach.

METHOD

In this section, we outline the comprehensive methodology employed in our study, encompassing the key stages of Convolutional Neural Networks (CNNs) application, model training, Class Activation Mapping (CAM), dimensionality reduction, and the evaluation metric of Intersection over Union (IoU). The general view of the study can be seen in Figure 3.

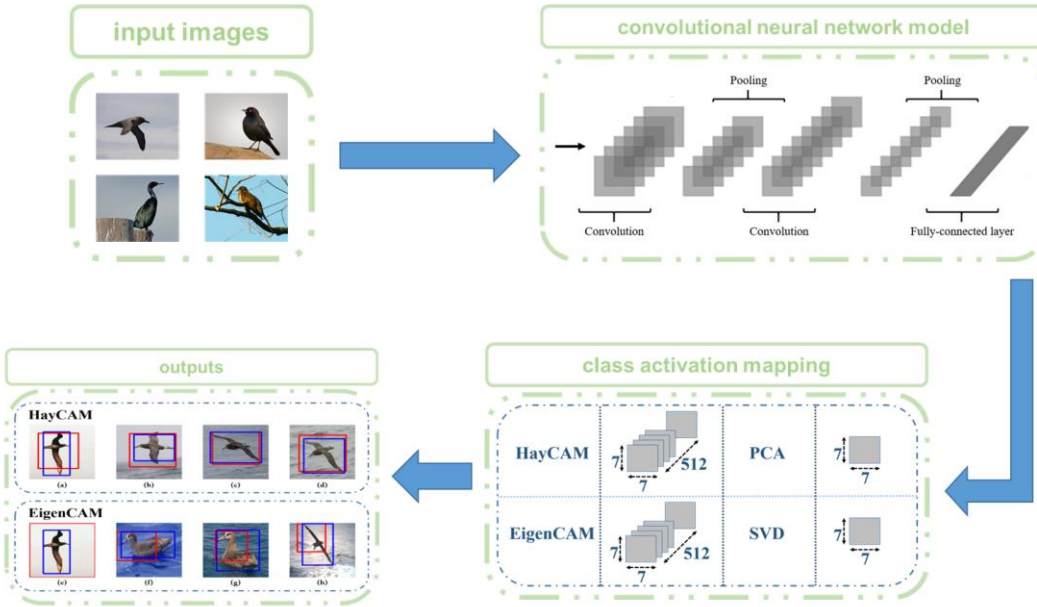

Figure 3. The General View of the Study

As shown in Figure 3, an input image is fed into a deep learning model, and class activation maps are obtained using the HayCAM and EigenCAM.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012), as a deep learning method, perform with high accuracy in applications such as image classification and object detection. A CNN model consists of two fundamental layers: the filtering layer and the classification layer as seen in Figure 4. In the filtering layer, convolution and pooling operations are conducted. The convolution operation is applied to the input image, resulting in a feature map.

Further convolution operations are applied to these obtained feature maps, extracting different features within the image. These features can be edges, textures, and more complex structures as the network deepens. The pooling operation reduces the dimensions of the obtained feature maps. For instance, applying a 2x2 pooling to a 256x256 filter yields a feature map of size 128x128.

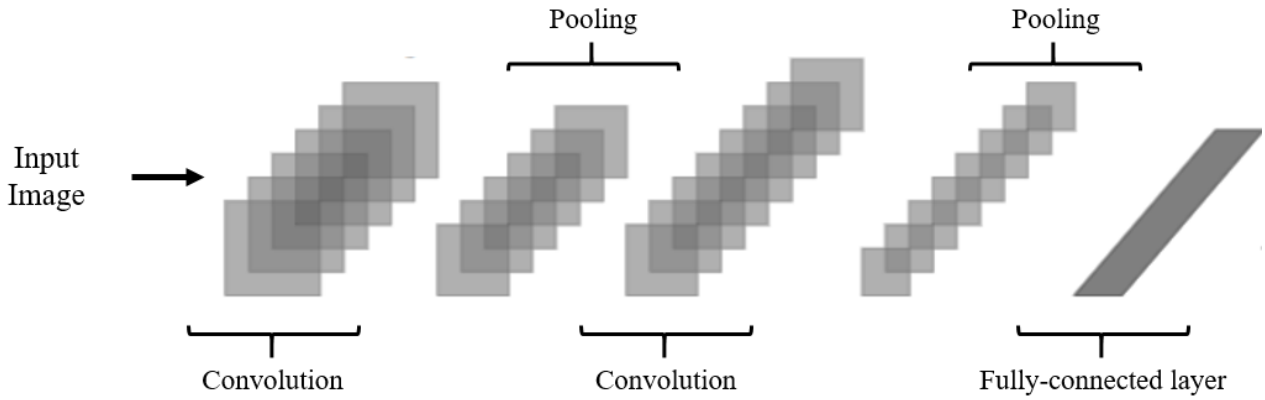

Figure 4. A CNN Model Including 2 Convolutions, 2 Poolings, and A Fully-Connected Layer

CNNs have emerged as a powerful class of neural network architectures, particularly for image and video analysis. CNNs automatically learn hierarchical representations of visual features directly from raw input data.

Fully connected layers, typically situated at the end of the network, produce final predictions. The trainable parameters of CNNs allow them to adapt to the data through the process of backpropagation, where the model learns optimal weights to minimize a predefined loss function.

Model Training

Training robust models often involves leveraging transfer learning, a technique that harnesses the knowledge gained from pre-trained models on large datasets and adapts it to new, specific tasks. One widely used architecture for such transfer learning tasks is ResNet-18, a variant of the Residual Network (ResNet) architecture (He et al., 2016).

Transfer learning is a technique that involves using a pre-trained model, which has been trained on a large dataset, and adapting it to a specific task or domain (Kornblith et al., 2019; Simonyan & Zisserman, 2014). This method is especially useful when there is limited labeled data available for the target task. The pre-trained model acts as a feature extractor, capturing general features from the original dataset. In the context of image classification, ResNet-18, with its ability to learn hierarchical representations, has proven to be an effective choice for transfer learning.

The training procedure for an AI model using transfer learning with ResNet-18 involves the following steps:

- Pre-training: Start with a pre-trained ResNet-18 model, often pre-trained on a large dataset like ImageNet, to capture general features.
- Feature Extraction: Freeze the weights of the early layers and extract features from the pre-trained model for the specific dataset of interest.
- Fine-tuning: Modify the last few layers of the network to match the target task's output, and fine-tune the model on the new dataset to adapt it to the specific classification or detection task.
- Training: Train the modified ResNet-18 model on the target dataset with labeled examples, using techniques like stochastic gradient descent or its variants.

This approach significantly accelerates training for specific tasks, especially when labeled data is limited, and it capitalizes on the ability of ResNet-18 to capture intricate hierarchical features. The combination of transfer learning and the ResNet-18 architecture has become a staple in various computer vision applications.

Class Activation Mapping

CAMs is a technique that enhances interpretability and localizes the areas within an image that contribute to the network's classification decision (B. Zhou et al., 2016). CAMs provide a visual representation of the importance of different regions in the input image, revealing where the network focuses its attention during the classification process. By incorporating CAMs into CNNs, we not only harness the network's ability to learn hierarchical features but also gain valuable insights into the discriminative regions contributing to its decisions. This fusion of CNNs and CAMs not only enhances accuracy in tasks like image classification and object detection but also facilitates a deeper understanding of the reasoning behind the network's predictions, making it a potent tool for XAI applications.

CAMs are an interpretability technique that provides visual insights into the decision-making process of neural networks, particularly in image classification tasks. CAMs highlight the regions of an input image that significantly contribute to the model's classification decision.

In its simplest form, a CAM is generated by taking the weighted sum of the feature maps from the final convolutional layer of a neural network. Let F_i be the activation map of the *i*-th channel in the final convolutional layer, and w_i be the corresponding weight for that channel. The CAM for a specific class *c* is computed as follows (Eq. 1):

$$\mathrm{CAM}_c = \sum_i w_i \cdot F_i \tag{1}$$

The weights w_i are derived from the final fully connected layer or global average pooling layer of the network.

Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) enhances the basic CAM (B. Zhou et al.., 2016) concept by incorporating gradient information. It computes the gradients of the predicted class score with respect to the feature maps. The importance weights w_i are proportional to the gradients, and the CAM is calculated similarly. Grad-CAM provides finer localization of discriminative regions.

EigenCAM (Muhammad & Yeasin, 2020) extends the CAM concept by introducing an eigen decomposition step. It involves computing the singular values of the weight matrix used in the CAM computation by using SVD. EigenCAM aims to capture more nuanced information about the significant features contributing to the classification decision.

HayCAM (A. Ornek & Ceylan, 2022) further refines the CAM approach by incorporating PCA. This allows for a more detailed representation of the relationships among features, leading to enhanced localization accuracy compared to traditional CAM methods. Even if HayCAM reduces the number of filters to 19, it was arranged to reduce the number of filters to 1 filter.

CAMs, including Grad-CAM, EigenCAM, and HayCAM, have found applications in various domains, such as object detection, image segmentation, and weakly-supervised object detection. Comparative analysis of these techniques, considering factors like interpretability, localization accuracy, and computational efficiency, helps in selecting the most suitable CAM variant.

Dimensity Reduction

Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are two powerful mathematical techniques that have found widespread application in a variety of fields, including data science, signal processing, computer vision, and machine learning. These techniques are used for dimensionality reduction and feature extraction, which are essential tasks in many applications where high-dimensional data needs to be processed and analyzed efficiently.

PCA is a technique that aims to transform the original data into a new coordinate system, where the maximum variance lies along the first principal component, the second maximum along the second component, and so forth. The transformation is achieved through the process of identifying the eigenvectors and eigenvalues of the covariance matrix of the data. This involves the use of statistical techniques to analyze the data and identify patterns and trends that can be used to create new insights and understandings.

The covariance matrix, denoted as *C*, is calculated as follows (Eq. 2):

$$C = \frac{1}{k-1}\sum_{i=1}^{k}(B_i - \bar{B})(B_i - \bar{B})^T \tag{2}$$

where B_i represents the data points, $\bar{B}$ is the mean of the data, and *k* is the number of data points.

The eigenvectors of *C* form the principal components, and the corresponding eigenvalues indicate the variance along each component.

SVD is another powerful mathematical technique that is closely related to PCA. It is used to decompose a matrix into its constituent parts, which can be used for a variety of tasks, including image compression, data compression, and signal processing. SVD decomposes a matrix *B* into three other matrices: *Y*, Σ, and P^T. For a given matrix B_{mxk}, where *m* is the number of rows and *k* is the number of columns (Eq. 3):

$$B = Y\Sigma P^T \tag{3}$$

where:

- *Y* is a matrix that contains the left singular vectors and is orthogonal,
- Σ is a diagonal matrix that has singular values on its diagonal,
- P^T is the transpose of a matrix that contains the right singular vectors and is orthogonal.

The singular values in Σ represent the square roots of the eigenvalues of B^TB or BB^T.

While both PCA and SVD are closely related, with PCA essentially being a specific application of SVD, they differ in their objectives. PCA focuses on finding the axes of maximum variance, whereas SVD decomposes a matrix into its singular values and vectors. SVD provides a full decomposition, making it more versatile, but for dimensionality reduction, PCA is often more intuitive and computationally efficient.

Intersection over Union (IoU) in Weakly-Supervised Object Detection

Weakly-supervised object detection (Shao et al., 2022) leverages methods such as CAM to overcome the challenge of limited labeled data. The activation maps are employed to guide the learning process, enabling the network to identify objects without requiring precise bounding box annotations. This approach significantly reduces the manual labeling effort while still achieving competitive detection performance. By harnessing weak supervision, CAM-based object detection models can learn to recognize objects with only image-level labels, demonstrating promising potential for deployment across various domains where labeled data is scarce or expensive to obtain.

Intersection over Union (IoU) is a metric that is widely used to evaluate the accuracy of object detection models. The IoU provides a quantitative measure of how well the predictions of the bounding boxes align with the ground truth bounding boxes. The IoU is calculated by dividing the area of overlap between the predicted and ground truth bounding boxes by the area of their union. A bounding box, intersection, and union areas are shown in Figure 5. A high IoU score indicates that the model's predictions are accurate and that the bounding boxes are tightly aligned with the ground truth, while a low IoU score suggests that the model's predictions are inaccurate and require further refinement.

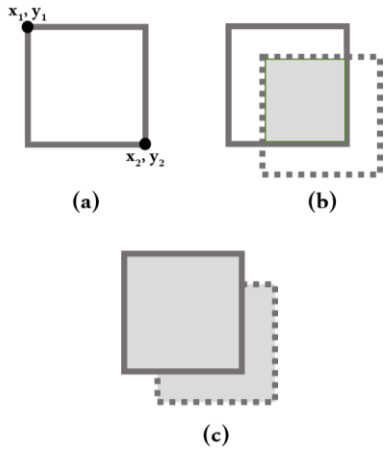

Figure 5. Bounding Box, Intersection, and Union Regions

Let BB_{pred} and BB_{gt} represent the predicted and ground truth bounding boxes, respectively. The area of overlap $Area_{\text{overlap}}$ is given by the intersection of the two bounding boxes (Eq. 4, Eq. 5, Eq. 6):

$$a_1 = \max\left(0, \min\left(x_{\max}^{\text{pred}}, x_{\max}^{\text{gt}}\right) - \max\left(x_{\min}^{\text{pred}}, x_{\min}^{\text{gt}}\right)\right) \tag{4}$$

$$a_2 = \max\left(0, \min\left(y_{\max}^{\text{pred}}, y_{\max}^{\text{gt}}\right) - \max\left(y_{\min}^{\text{pred}}, y_{\min}^{\text{gt}}\right)\right) \tag{5}$$

$$Area_{\text{overlap}} = a_1 \times a_2 \tag{6}$$

Where x_{min}, x_{max}, y_{min}, and y_{max} denote the minimum and maximum coordinates of the bounding boxes.

The area of union $Area_{\text{union}}$ is calculated as the sum of the individual areas of the predicted and ground truth bounding boxes minus the area of overlap (Eq. 7):

$$Area_{\text{union}} = Area_{\text{pred}} + Area_{\text{gt}} - Area_{\text{overlap}} \tag{7}$$

The IoU is then computed as the ratio of the area of overlap to the area of union (Eq. 8):

$$IoU = \frac{Area_{\text{overlap}}}{Area_{\text{union}}} \tag{8}$$

In weakly-supervised object detection, IoU metrics play a crucial role in assessing the localization accuracy of detected objects. The IoU metric becomes particularly valuable in scenarios where object boundaries are not precisely defined, allowing for a more flexible evaluation of detection performance.

EXPERIMENTS AND RESULTS

The study was conducted to assess the impact of object size on object detection performance. To achieve this, the ratios of objects within images were calculated based on the corresponding image areas. The calculation was carried out in intervals such as 100% to 90%, 90% to 80%, and so forth, down to 10% to 0%. This method allowed for a systematic exploration of how varying object sizes influence the object detection capabilities of the model. The results from EigenCAM and HayCAM, as presented in Table 2, provide a detailed insight into the IoU values across different object/image ratios.

Table 2. IoU Results of Detections

Object/Image Ratio	EigenCAM	HayCAM
100% - 90%	**0.5583**	0.5289
90% - 80%	0.4853	**0.5528**
80% - 70%	0.4218	**0.5536**
70% - 60%	0.3425	**0.5190**
60% - 50%	0.2674	**0.4574**
50% - 40%	0.2034	**0.3794**
40% - 30%	0.1684	**0.3116**
30% - 20%	0.1595	**0.2684**
20% - 10%	0.1732	**0.2047**

Table 2 shows the IoU results for two different methods, EigenCAM and HayCAM, for various object/image ratios. The IoU is a measure of the overlap between two boxes, and it is used to evaluate the performance of object detection algorithms. Table 2 shows that HayCAM outperforms EigenCAM for most of the object/image ratios. For example, for the object/image ratio of 80-70%, HayCAM achieves an IoU of 0.5536, while EigenCAM achieves an IoU of 0.4218. However, for the object/image ratio of 100-90%, EigenCAM achieves the highest IoU of 0.5583, while HayCAM achieves an IoU of 0.5289. EigenCAM method achieves an average 30.88% IoU, while the HayCAM method achieves 41.95% IoU. The visual results are given in Figure 6.

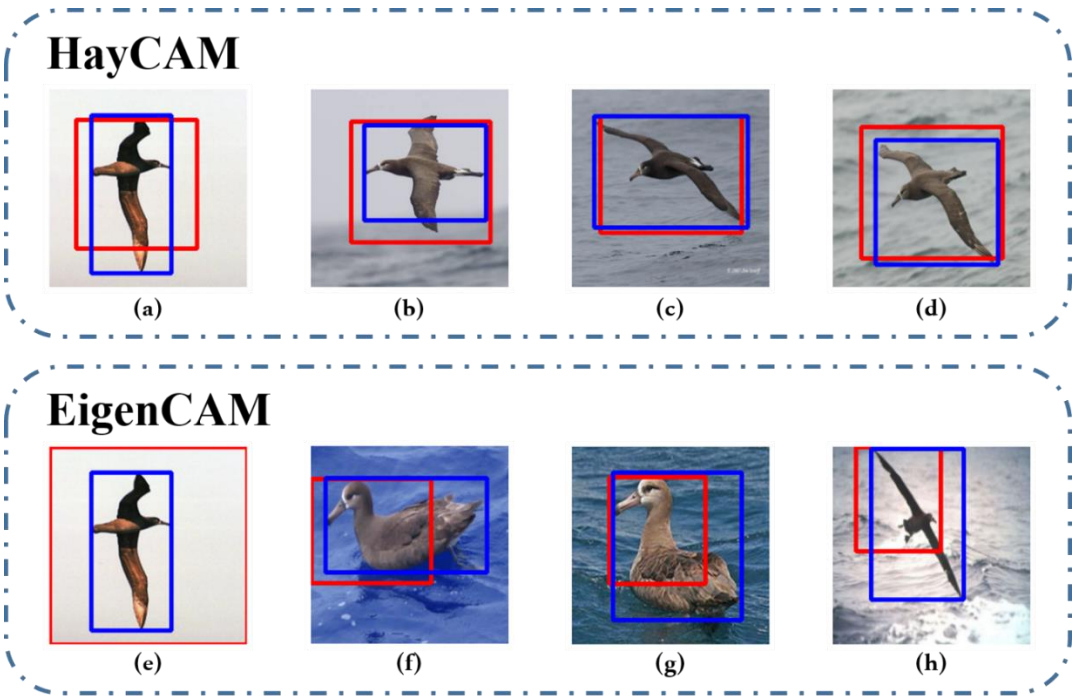

Figure 6. Visual Results of HayCAM (a, b, c, d) and EigenCAM (e, f, g, h). Object/Image Ratios Are (a) 32.06% (b) 28.54% (c) 44.40% (d) 42.51% (e) 32.06% (f) 38.28% (g) 51.06 (h) 35.74

The results suggest that HayCAM is a more effective method for object detection than EigenCAM, especially for object/image ratios below 80%. The results also show that the performance of both methods decreases as the object/image ratio decreases. This is expected, as it becomes more difficult to detect objects as they become smaller in relation to the image size. The results can be used to guide the selection of object detection methods for different applications, depending on the object/image ratio and the desired level of performance.

DISCUSSION

Table 2 shows the Intersection of Union (IoU) values for EigenCAM and HayCAM across various object/image ratios. This discussion delves into the observed trends, highlighting the strengths and weaknesses of each method in localizing objects within images.

Examining the table reveals a notable variation in performance between EigenCAM and HayCAM across different object/image ratios. EigenCAM exhibits a strong start in the 100% to 90% ratio, outperforming HayCAM with an IoU of 0.5583. However, a shift occurs in favor of HayCAM in subsequent intervals (90% to 80% and beyond), where it consistently achieves higher IoU values than EigenCAM. This trend suggests that HayCAM is particularly adept at handling diverse object sizes within images, showcasing its robustness in capturing relevant features.

One striking observation is the consistent outperformance of HayCAM throughout the majority of the presented ratios. In particular, 80% to 70% and 90% to 80% intervals stand out with HayCAM achieving IoU values of 0.5536 and 0.5528, respectively, demonstrating its capacity to accurately identify and localize objects across a spectrum of sizes. The sustained superiority of HayCAM implies its efficacy in scenarios where objects exhibit variations in scale, making it a promising choice for interpretable object detection models.

EigenCAM's initial strength, as evidenced by the high IoU in the first ratio, diminishes in subsequent intervals. The declining trend suggests that EigenCAM may face challenges in consistently localizing objects as their sizes vary. Understanding these nuances is crucial for practitioners to select interpretable methods based on the specific requirements of their object detection tasks.

The comparative analysis of EigenCAM and HayCAM results holds significant implications for model interpretability in object detection. While EigenCAM may excel in certain scenarios, HayCAM's consistent

performance across a broader range of object sizes highlights its versatility. This insight enables researchers and practitioners to make informed choices regarding the selection of interpretable methods based on the characteristics of their datasets and the challenges posed by varying object scales.

The discussed results underscore the importance of evaluating interpretable methods in the context of diverse object scales. The observed trends contribute valuable insights into the strengths and limitations of EigenCAM and HayCAM, guiding future endeavors in the development and application of interpretable models for object detection tasks.

CONCLUSION

In conclusion, the presented investigation into the object detection performance of EigenCAM and HayCAM across varying object/image ratios provides valuable insights into the interpretability and adaptability of these methods. The experimental results revealed nuanced trends, with EigenCAM showcasing initial strength in accurately localizing objects within the 100% to 90% ratio but experiencing a decline in performance in subsequent intervals. Conversely, HayCAM consistently outperformed EigenCAM, demonstrating robust localization capabilities across a spectrum of object sizes. This suggests that the choice between these interpretable methods should be contingent on the specific challenges posed by varying object scales within images.

The observed trends in the IoU values highlight the importance of understanding the intricacies of interpretable models for object detection. The consistent superiority of HayCAM implies its potential applicability in real-world scenarios where objects exhibit diverse scales, enhancing the model's ability to provide meaningful and reliable insights into decision-making processes. The findings contribute to the broader discourse on model interpretability, emphasizing the need for nuanced assessments in diverse contexts.

For future work, it is crucial to delve deeper into the factors influencing the performance of HayCAM. Conducting a thorough analysis of the model's response to specific object characteristics and dataset attributes can provide a more nuanced understanding. Additionally, exploring enhancements to these CAM methods or combining them with other interpretability techniques may further improve their robustness and generalizability. The integration of additional datasets with distinct characteristics can contribute to a more comprehensive evaluation of these interpretable methods in various real-world scenarios. Addressing these aspects will pave the way for the refinement and advancement of interpretable models, fostering their broader adoption in complex object detection tasks.

ACKNOWLEDGMENT

This study was supported by Huawei Türkiye R&D Center.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., & Barbado, A. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 839–847). https://doi.org/10.1109/WACV.2018.00097

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the ieee conference on computer vision and pattern recognition (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better imagenet models transfer better? In Proceedings of the ieee conference on computer vision and pattern recognition (pp. 2661–2671). https://doi.org/10.1109/CVPR.2019.00277

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25. https://doi.org/10.1145/3065386

Muhammad, M. B., & Yeasin, M. (2020). Eigen-cam: Class activation map using principal components. In 2020 international joint conference on neural networks (ijcnn) (pp. 1–7). https://doi.org/10.1109/IJCNN48605.2020.9206626

Ornek. (2023). Developing a new explainable artificial intelligence method (doctoral dissertation). Konya Technical University. (No DOI available for the dissertation)

Ornek, A., & Ceylan, M. (2022). Haycam: A novel visual explanation for deep convolutional neural networks. Traitement Du Signal, 39 (5), 1711–1719. https://doi.org/10.18280/ts.390529

Ornek, A. H., & Ceylan, M. (2022). A novel approach for visualization of class activation maps with reduced dimensions. In 2022 innovations in intelligent systems and applications conference (asyu) (pp. 1–5). https://doi.org/10.1109/ASYU56188.2022.9925400

Ornek, A. H., & Ceylan, M. (2023). Codcam: A new ensemble visual explanation for classification of medical thermal images. Quantitative InfraRed Thermography Journal, 1–25. https://doi.org/10.1080/17686733.2023.2167459

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the ieee international conference on computer vision (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Shao, F., Chen, L., Shao, J., Ji, W., Xiao, S., Ye, L., Xiao, J. (2022). Deep learning for weakly-supervised object detection and localization: A survey. Neurocomputing. https://doi.org/10.48550/arXiv.2105.12694

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Stewart, G. W. (1993). On the early history of the singular value decomposition. SIAM review, 35 (4), 551–566. https://doi.org/10.1137/1035134

van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis, 79 , 102470. doi: https://doi.org/10.1016/j.media.2022.102470

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). Caltech (Tech. Rep. No. CNS-TR-2011-001). California Institute of Technology. (Technical reports typically do not have DOIs)

Wu, S. X., Wai, H.-T., Li, L., & Scaglione, A. (2018). A review of distributed algorithms for principal component analysis. Proceedings of the IEEE, 106 (8), 1321–1340. https://doi.org/10.1109/JPROC.2018.2846568

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the ieee conference on computer vision and pattern recognition (pp. 2921–2929). https://doi.org/10.1109/CVPR.2016.319

Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., & Yang, R. (2019). Iou loss for 2d/3d object detection. In 2019 international conference on 3d vision (3dv) (pp. 85–94). https://doi.org/10.1109/3DV.2019.00019