M-CAM: Visual Explanation of Challenging Conditioned Dataset with Bias-reducing Memory

Seongyeop Kim
seongyeop@kaist.ac.kr
Yong Man Ro
ymro@kaist.ac.kr

Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Korea

Abstract

We introduce a framework that enhances visual explanation of class activation map (CAM) with key-value memory structure for deep networks. We reveal challenging conditions inherently existing in several datasets that degrade the visual explanation quality of existing CAM-based visual explanation methods (e.g. imbalanced data, multi-object co-occurrence) and try to solve it with the proposed framework. The proposed Bias-reducing memory module learns spatial feature representation of different classes from trained networks and stores each different semantic information in separate memory slots, while it does not require any modification to the existing networks. Furthermore, we propose a novel visual explanation method accompanied by a memory slot searching algorithm to retrieve semantically relevant spatial feature representation from the memory module and make visual explanation of network decisions. We evaluate our visual explanation framework with datasets of challenging conditions including several medical image datasets and multi-label classification datasets. We qualitatively and quantitatively compare it with existing CAM-based methods to demonstrate the strength of our framework.

1 Introduction

Explainable artificial intelligence (XAI), for computer vision in specific, comprises a wide range of research opportunities to make explanations of deep network decisions with curiosities over hidden operations proceeding inside the deep networks. From intrinsic to post-hoc methods and from local to global explanation methods, numerous researches of quality have been published to explain decision making process of deep networks. In this paper, we specify a well-known visual explanation tool for deep Convolutional Neural Network (CNN), class activation map (CAM), and try to overcome the limits of it and its variations.

CAM [20] marks local areas of an image, that have positive influence on deep networks decision making, with upsampled feature maps with respective importance weights. It takes advantage of internal components of deep networks such as gradients and feature maps to generate visual explanations. Grad-CAM [14] and Grad-CAM++ [1] are generalizations of CAM, and several variations of CAM have been proposed [2, 3, 11, 13, 16]. However,

the quality of visual explanation obtained by CAM-based methods may vary depending on the training environment of the target deep network. That is, challenging conditions such as imbalanced distribution of class, lack of training samples, or frequent co-occurrence of multiple objects in training dataset may not guarantee the reliability of internal components of deep networks leading to degradation of credibility on generated visual explanations.

In this paper, we raise two major types of problematic biases of deep network that may hinder the quality of visual explanation. First is the problem of co-occurrence in a multi-label classification environment. In a multi-label classification dataset, specific multiple classes appearing in a single image is a frequent event (e.g. horse and person). Disentangling such co-occurring classes while generating visual explanation would be the first goal. The second problem is about the imbalanced class distribution of training dataset and weak decision boundaries of deep network that consequently follow. This problem is commonly found with medical image dataset even being accompanied with insufficient number of training samples for specific classes. Enhancement of visual explanation quality for the insufficiently represented classes would be the next objective.

To tackle such problems, we propose Bias-reducing memory module that provides quality visual explanations even with datasets of challenging conditions. To complement biases of the target network caused by the inherent challenging conditions of the training dataset, the proposed key-value structure memory module learns the distribution of spatial feature representation from the target deep network and discretely organizes the distributions into separate memory slots. The memory module does not require specific CNN structure and the proposed learning scheme allows the memory module to be trained by itself without any modification on the target deep network. Then we take advantage of the memory module to acquire quality visual explanation in two steps, memory slot searching with feature perturbation and adjustment of importance weights on the feature activation maps.

We summarize the contributions of the proposed method into three. First, we propose a key-value structure Bias-reducing memory module and its learning scheme to organize the spatial feature distributions of dataset into discrete memory slots. Second, we introduce a new CAM-based visual explanation method with the memory module, M-CAM, that provides solid visual explanation even in challenging training conditions. Third, to verify the strength of M-CAM in challenging conditions, we conduct experiments on four medical image datasets [4, 6, 12, 17] and MS COCO [9], comparing our method with the existing CAM-based visual explanation methods.

2 Related Work

2.1 Class Activation Map

The concept of class activation map (CAM) is first introduced by [20]. This first work aggregates the feature maps obtained from the last convolutional layer, with the importance weight assigned to each feature map, to generate activation map that has positive influence on the decision made by the CNN. Afterward, Grad-CAM [14] generalizes the concept of CAM, verifying that generation of activation map does not require specific structure (global average pool layer at the end of the network) of CNN anymore. Grad-CAM++ [1] further generalizes Grad-CAM, enhancing its object localization by taking advantage of higher order derivatives for importance weight assignment. Several variations of CAM have been published as well. Score-CAM [16] generates mask from each feature map and utilizes the prediction score

of masked images as the importance weights. U-CAM [13] brings the Bayesian nature of predicting uncertainty of the model and the dataset itself [5] and adopted it to prune out uncertain region in the visual explanation. Eigen-CAM [11] takes advantage of the principle components of feature representations to provide visual explanation independent of classification layers. Axiom-based Grad-CAM [3] introduces sensitivity and conservation as two new axioms to evaluate CAM-based visual explanation methods. Ablation-CAM [7] presents a gradient-free methodology to generate visual explanation. We point out that the problems we tackle regarding inherent challenging conditions of the dataset have not been thoroughly discussed yet, and we try to solve such drawbacks of CAM-based methods with the proposed methodology.

2.2 Memory Network

External memory network [15, 18] is introduced to augment neural network, and it functions to store useful information acquired by the neural network. Depending on the task the target network is designed for, the information stored in the memory can be utilized in varied ways. One of the structures of the memory network is key-value structure [10]. It stores diverse patterns of feature representation in the key memory, and utilize it as a blueprint to make inference for the target information which is stored in the value memory. As a simple example in computer vision, key memory stores feature representation of trained images while the value memory stores class-related semantic information to output class prediction for the input image. We design the internal structure of the key-value memory inspired by the concept of sparse dictionary learning [2], where diverse information can be stored sparsely over different memory slots so that one can utilize the memory network to the most extent. We design the memory network module to be applied any type of CNN structure and the learning scheme allows the memory network module to learn useful information by itself.

3 Bias-reducing Memory

In this section, we describe the overall structure of Bias-reducing memory module and which information we would like to learn from the target network. Then we describe the objective functions we design to train the proposed memory module. The proposed objective functions effectively guide the memory module to learn desired information while not affecting the target deep network and its performance.

3.1 Overall Memory Concept

End-to-end structure of memory network in machine learning [15] is first introduced by Sukhbaatar et al. for question answering task. Key-value memory [10] is a generalization of the traditional memory network, allowing it to flexibly store feature representation distribution as a prior knowledge and to take advantage of it with complex transform enabled by the memory.

Application of key-value memory involves two major steps, which are key addressing and value reading. Given an embedded query value $q \in \mathbb{R}^c$, similarity between q and each slot of key memory $K_i \in \mathbb{R}^c$ is measured. An address vector $p \in \mathbb{R}^{1\times N}$ is obtained for a key memory K with N slots, where each scalar value of p represents similarity between the query and each memory slot:

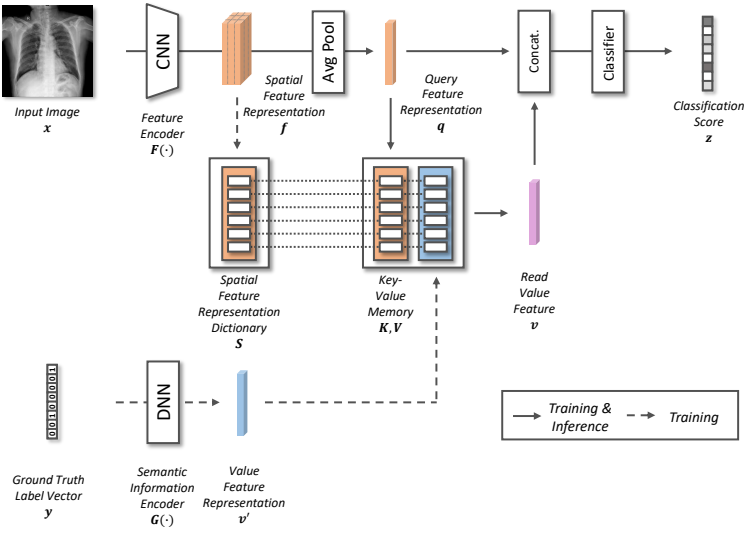

Figure 1: Overall structure of the memory module being applied to the target network.

$$p_i = Softmax(\frac{q \cdot K_i}{\|q\| \, \|K_i\|}) \tag{1}$$

where $i = 1, 2, \ldots, N$ and $Softmax(z_i) = e_i^z / \sum_{j=1}^{N} e_j^z$.

In value reading step, the value memory is accessed by the key address vector p as a set of relative weights of importance for each slot. The read value $v \in \mathbb{R}^c$ is obtained such that $v = pV$, where $V \in \mathbb{R}^{N \times c}$ is a trained value memory with N slots. By doing so, key-value memory structure allows it to flexibly access to desired information stored in the value memory corresponding to different query values.

3.2 Application of the Proposed Memory Module

Figure 1 describes an overall flow on how the proposed Bias-reducing memory module learns desired information from the target network. Given a pre-trained feature encoder F of the target network, the memory module takes the spatial feature representation $f \in \mathbb{R}^{w \times h \times c}$, query feature representation $q \in \mathbb{R}^c$ and a value feature representation $v' \in \mathbb{R}^c$ as input for training. We devise G to map the hot encoded ground truth label vector y into the same number of dimensionality as q. In the inference step, f and v' are not required. In both training and inference step, the memory module outputs read value feature $v \in \mathbb{R}^c$ as an output, and the classifier takes a concatenated vector of q and v as an input to output classification score z.

Figure 2 describes details of the memory module. Key address vector p is obtained as same as Eq. 1 in section 3.1., and the spatial feature address vector p_s and value address vector p' is obtained in the same manner. $K_i \in \mathbb{R}^c$ represents the query information stored in the i th slot of the key memory K ($i = 1, 2, \ldots, N$) to retrieve semantic information from the value memory V, $V_i \in \mathbb{R}^c$ represents the semantic information stored in the i th slot of V and $S_i \in \mathbb{R}^{w \times h \times c}$ represents the distribution of spatial feature representation stored in the i th slot of S, spatial feature representation dictionary. We guide the memory module to store

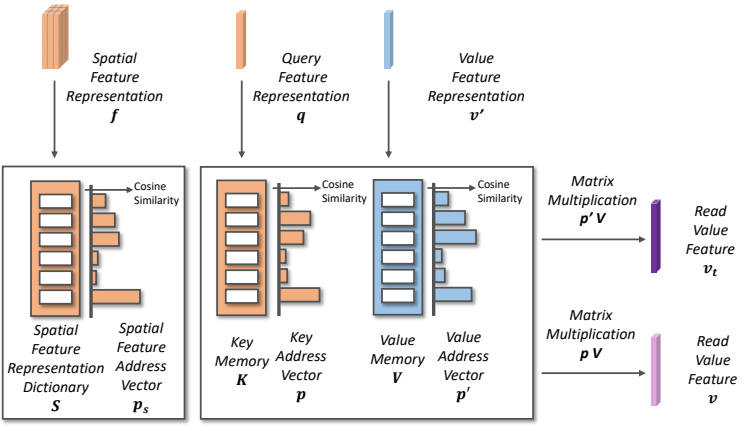

Figure 2: Details of the memory module in training phase.

corresponding information at the same sequential location of slot. In other words, if the second slot of V turns out to contain semantic information related to *dog* class, we guide the second slot of S to learn corresponding distribution of spatial feature representation of *dog* class. We explain the objective functions and its roles in the following subsection. In the training step, we obtain two read value features v_t and v where v_t is only used for training the memory module. We obtain the read value feature $v = pV$ as described in the preceding subsection, while v_t is obtained by a matrix multiplication of V and value address vector p' as relative importance weights for memory slots.

3.3 Training Memory Module

To effectively guide Bias-reducing memory module to learn the distribution of spatial feature representation with the corresponding semantic information distilled from the target network, we design three objective functions $\mathcal{L}_{classifier}$, $\mathcal{L}_{sparse}$, and $\mathcal{L}_{address}$.

As in Figure 1, a new classifier has to be trained from the scratch in order to train the memory module. We devise $\mathcal{L}_{classifier}$ as,

$$\mathcal{L}_{classifier} = BCE(fc(cat(v_t, f)), Y) + BCE(fc(cat(v, f)), Y), \tag{2}$$

where $BCE(y, \hat{y}) = -\frac{1}{N}\Sigma_1^N y_i \cdot log(\hat{y}_i) + (1-y_i) \cdot log(1-\hat{y}_i)$ is Binary Cross Entropy loss function, $fc(\cdot)$ is a fully connected layer classifier, and $cat(\cdot)$ represents concatenation between two vectors. Since v_t and v are obtained from the value address vector p' and the key address vector p respectively, each term of $\mathcal{L}_{classifier}$ is devised to train the value memory V and key memory K each.

We want the value memory V to store semantic information encoded by G, and expect the memory module to output the read value feature v as similar as encoded value feature v' even in the inference phase. While $\mathcal{L}_{sparse}$ being applied for training the memory module, sparse representations of semantic information are learned over the memory slots and the memory module forms a linear combination of each slot to output the read value feature v. We devise $\mathcal{L}_{sparse}$ as L2 norm between the two read value features v_t and v:

$$\mathcal{L}_{sparse} = \frac{1}{N}\sum_{i=1}^{N}(v_i - v_{t_i})^2. \tag{3}$$

To jointly store corresponding information at the same sequential location of the memory slots at S, K, and V, we devise an address matching objective function $\mathcal{L}_{address}$. To effectively trace back the spatial feature representation distribution of specific class from the corresponding semantic information, $\mathcal{L}_{address}$ guides the spatial feature representation dictionary and key memory to output similar address vectors p_s and p to the value address vector p'. $\mathcal{L}_{address}$ is as follows,

$$\mathcal{L}_{address} = KL(p' \parallel p_s) + KL(p' \parallel p), \tag{4}$$

where $KL(p \parallel q) = -\Sigma_1^N p_i \cdot log(q_i/p_i)$ is Kullback-Leibler divergence [8]. We sum the three of the introduced objective functions to train the memory module (S, K, and V), a classifier, and the semantic information encoder G while the feature encoder F remains fixed. Hence the final objective function is

$$\mathcal{L} = \mathcal{L}_{classifier} + \mathcal{L}_{sparse} + \mathcal{L}_{address}. \tag{5}$$

4 M-CAM: Generating Visual Explanation

In this section we describe a slot searching algorithm to trace back the spatial feature representation distribution of a desired class and how we utilize the obtained distribution to generate quality visual explanation by the importance weight adjustment procedure.

4.1 Slot Searching Algorithm by Feature Perturbation

We devise a slot searching algorithm to disclose which slot contains information that is most closely related to the desired class of our interests. The intuition of the algorithm is that, we want to observe the prediction score decrease while each slot of the memory module is perturbed with a random noise. Therefore we assume that a particular slot replaced with a noise leading to the highest score drop has the most closely related information to the target class $\hat{c}$. As described in Algorithm 1, given the query feature representation q_x of an input image x, a target class $\hat{c}$, and the original prediction score z of x, the algorithm returns the slot sequence number $n_{\hat{c}}$ that contains the most closely related information for the target class $\hat{c}$ in the trained memory module.

4.2 Importance Weight Adjustment by Memory

After we finish searching slot for the target class, we retrieve the corresponding distribution of spatial feature representation $S_{n_{\hat{c}}} \in \mathbb{R}^{w \times h \times c}$ from the $n_{\hat{c}}$ th slot of the dictionary S. Given a spatial feature representation f_x of an input image x, we measure cosine similarity between f_x and $S_{n_{\hat{c}}}$ in a channel-wise manner. Then we get a set of $\tau = \{\tau_1, \tau_2, \ldots, \tau_c\}$ where

$$\tau_i = \frac{f_{x_i} \cdot (S_{n_{\hat{c}}})_i}{\left\| f_{x_i} \right\| \left\| (S_{n_{\hat{c}}})_i \right\|}. \tag{6}$$

$f_{x_i} \in \mathbb{R}^{w \times h}$ and $(S_{n_{\hat{c}}})_i \in \mathbb{R}^{w \times h}$ are the activation map at the i th channel obtained from f_x and $S_{n_{\hat{c}}}$ respectively. τ_i represents how similar the spatial feature representation of the input image x at the channel i is to the one retrieved from the dictionary S. Here we assign importance weight w_i to each spatial feature representation map f_{x_i},

Algorithm 1 Slot Searching Algorithm by Feature Perturbation

Inputs: Query feature representation q_x of an input image x, a target class $\hat{c}$, and the original prediction score z of x.
Output: Slot sequence number $n_{\hat{c}}$ of the target class $\hat{c}$.

procedure SEARCH($q_x, \hat{c}, z_{\hat{c}}$)
  $n_{\hat{c}} \leftarrow 0$
  $max_{n_{\hat{c}}} \leftarrow 0$
  for $i = 1, 2, ..., N$ **do** ▷ Number of slots of each memory is N
    $K_{temp} \leftarrow K$ ▷ Make a copy of the key memory every iteration
    $\boldsymbol{w} \sim \mathcal{N}(0,1)$ ▷ Sample noise vector $\boldsymbol{w} \in \mathbb{R}^c$ from normal distribution
    $K_{temp_i} \leftarrow \boldsymbol{w}$ ▷ Perturb slot number n
    for $j = 1, 2, ..., N$ **do**
      $p_i \leftarrow Softmax(\frac{q_x \cdot K_{temp_j}}{\|q_x\| \|K_{temp_j}\|})$ ▷ Compute address vector
    end for
    $v \leftarrow p \cdot V$ ▷ Get read value feature
    $z' = fc(cat(v, f))$ ▷ Get a new class prediction score
    if $max_{n_{\hat{c}}} < z_{\hat{c}} - z'_{\hat{c}}$ **then**
      $max_{n_{\hat{c}}} \leftarrow z_{\hat{c}} - z'_{\hat{c}}$ ▷ Update the highest score decrease for class $\hat{c}$
      $n_{\hat{c}} \leftarrow i$ ▷ Update the sequence number of the slot
    end if
  end for
  return $n_{\hat{c}}$
end procedure

$$w_i = \sum_u \sum_v \frac{\partial z_{\hat{c}}}{\partial (f_{x_i})_{uv}} \cdot exp(\tau_i), \tag{7}$$

where the preceding term is the gradient propagating from the target class prediction score $z_{\hat{c}}$ to the feature representation map f_{x_i} with Euclidean coordinate u and v. We take exponential function on τ_i to map the output range of cosine similarity [-1,1] to positive number of range $[e^{-1}, e]$ giving more emphasis on the cosine similarity value that is close to 1. The intuition of the importance weight adjustment utilizing the memory module is that, we want to prune out spatial feature representations that are irrelevant to the target class $\hat{c}$ while giving more emphasis on the ones similar to the retrieved feature distribution $S_{n_{\hat{c}}}$. By taking weighted sum of f_{x_i} with the set of importance weight $w = \{w_1, w_2, \ldots, w_c\}$ over c channels, we generate the class activation map for visual explanation.

5 Experiments

5.1 Datasets

We select datasets that inherently include challenging conditions such as co-occurrence of multiple objects in a single image, class imbalance of training dataset, and lack of training dataset. MS COCO (COCO) [9] includes a large number of images with multiple objects appearing in a single image. In addition, we select COCO to verify the generalizability

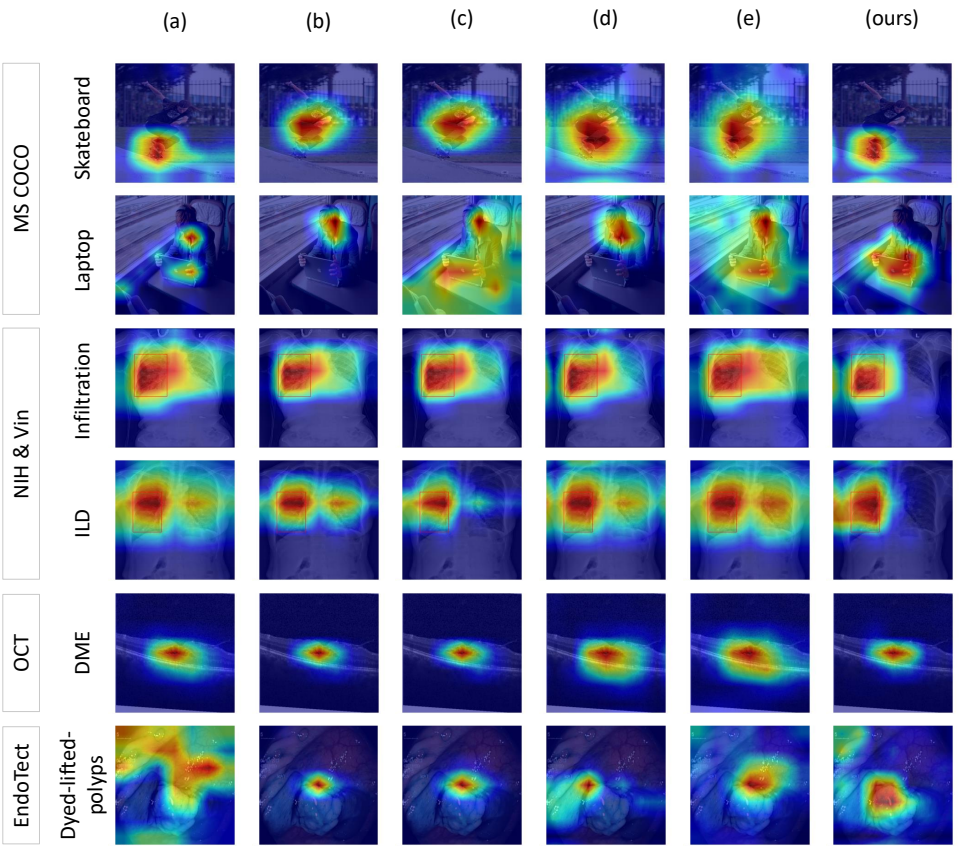

Figure 3: Qualitative results of M-CAM visual explanation over five experimental datasets. Dataset and the class information of the corresponding images is denoted on the left. (a): Ablation CAM [2], (b): EigenCAM [11], (c): EigenGradCAM [11], (d): GradCAM [14], (e): GradCAM++ [1], (ours): M-CAM.

of our method. Further, we select four medical image dataset for verification. NIH Chest X-ray 14 (NIH) [17] and VinDr-CXR (Vin) [12] are the frontal-view X-ray images where co-occurrence of multiple thorax diseases is commonly observed. Retinal optical coherence tomography (OCT) dataset [6] contains about 80K retina cross-section images with high class imbalance of retinal diseases. EndoTect Challenge dataset (Endo) [4] includes 10K endoscope images of human digestive system. With a small number of dataset samples, highly imbalanced class distribution of 23 classes is the challenging point of the dataset.

5.2 Qualitative Results Analysis

Figure 3 shows representative experimental results that reflect the strength of the proposed visual explanation framework. If training dataset (COCO) includes images of multi-object/class co-occurrence which induce bias to the deep network, the visual explanation of the existing CAM-based methods tend to highlight the context object (person) and the target obejct (skateboard and laptop) at the same time. The visual explanation results of the proposed framework highlight the context information being suppressed while giving emphasis on the

Dataset	Grad[14]	Grad++ [1]	Eigen[11]	Eigengrad[11]	Ablation [2]	Ours
NIH	36.44	39.28	40.20	39.37	**28.52**	39.25
Vin	33.93	27.35	26.56	**26.27**	28.09	28.19
COCO	69.06	50.28	49.44	**48.38**	49.65	68.82
OCT	43.84	49.98	43.13	45.39	49.43	**31.81**
Endo	85.62	86.71	85.50	84.17	86.13	**82.22**

Table 1: Performance measurement of six different visual explanation methods on the five datasets on Average Drop Percentage (lower is better) metric. The top performance for each dataset is written in bold.

Dataset	Grad[14]	Grad++ [1]	Eigen[11]	Eigengrad[11]	Ablation [2]	Ours
NIH	37.79	38.02	35.71	36.18	35.71	**38.25**
Vin	34.21	52.96	52.83	**53.31**	52.00	50.62
COCO	17.34	13.17	15.42	14.95	15.46	**17.39**
OCT	17.18	25.00	24.59	25.10	25.00	**32.21**
Endo	2.13	2.22	2.87	**2.96**	2.78	2.41

Table 2: Performance measurement of six different visual explanation methods on the five datasets on Percentage Increase in Confidence (higher is better) metric. The top performance for each dataset is written in bold.

target object solely.

In case of the chest X-ray images, multiple numbers of thoracic diseases are often detected in a single case, and the boundary of diseases are frequently overlapped or ambiguous. Hence the visual explanation of chest X-ray images often include irrelevant regions due to the ambiguity. With the bounding-box information provided, the heatmap of the proposed framework is densely distributed inside the box while the others highlight uncertain regions concurrently.

The OCT images contain relatively easy lesion boundaries so that the existing CAM-based methods highlight the lesion with a good quality as well, however, the proposed framework tends to spot concentrated region with the heatmap not spread out. EndoTect dataset is the most challenging dataset to train the network because of its less training samples and class imbalance, therefore the majority of the CAM-based methods had trouble providing visual explanation with good quality. However, we observe a wide area inside the lesion boundary has been covered by the generated heatmap of the proposed method.

5.3 Quantitative Experiments

To verify the advantage of the proposed method in a quantitative manner, we select four evaluation metrics for visual explanation, Average Drop Percentage, Percentage Increase in Confidence [1], Infidelity, and Sensitivity [19]. We select Average Drop Percentage and Percentage Increase in Confidence to evaluate how well the generated visual explanations highlight decisive region of images. Infidelity and Sensitivity are the objective metrics that evaluate the robustness of the explanation. Table 1 and 2 show that the proposed method outperformed the five existing CAM-based methods on two and three datasets in Average Drop Percentage and Percentage Increase in Confidence metric each, showing promising

Dataset	Grad++ [1]	Eigengrad[11]	Ablation [2]	Ours
NIH	0.4943	0.3267	0.6664	**0.0739**
Vin	0.6686	0.5934	**0.2173**	0.4649
COCO	0.3733	0.1396	0.2675	**0.0707**
OCT	0.3564	0.2899	0.3312	**0.0058**
Endo	0.0259	**0.0176**	0.0243	0.5980

Table 3: Performance measurement of four different visual explanation methods on the five datasets on Infidelity (lower is better) metric. The top performance for each dataset is written in bold.

Dataset	Grad++ [1]	Eigengrad[11]	Ablation [2]	Ours
NIH	0.0401	0.0279	0.1406	**0.0069**
Vin	0.0961	0.0579	0.1247	**0.0255**
COCO	0.0924	0.0763	0.1503	**0.0579**
OCT	0.0711	0.0816	0.1939	**0.0084**
Endo	**0.0609**	0.1289	0.0736	0.0771

Table 4: Performance measurement of four different visual explanation methods on the five datasets on Sensitivity (lower is better) metric. The top performance for each dataset is written in bold.

results on the rest of datasets. Table 3 and 4 show strong robustness of the proposed method compared with the three recent CAM-based methods. Experimental settings and details regarding the evaluation metrics are described in the supplementary material.

6 Conclusion

In this paper, we propose a new framework to provide class activation map-based visual explanation for datasets with challenging conditions causing bias to the deep network. We devise Bias-reducing memory to discretely store the distribution of spatial feature representation in different slots of the memory with the corresponding semantic information from the target deep network. With the slot searching algorithm by feature perturbation, we trace back which semantic information is stored in each memory slot and further utilize the retrieved distribution of spatial feature representation to enhance the quality of the class activation map. With the experiments done over four medical image datasets and MS COCO that inherently contain challenging conditions causing bias, we verify the strength of the proposed framework in such environment by comparative experiments with the existing visual explanation methods.

Acknowledgements

This work was partly supported by the IITP grant funded by the MSIT (No. 2020-0-00004).

References

[1] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[2] Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 972–980, 2020. doi: 10.1109/WACV45572.2020.9093360.

[3] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. In *British Machine Vision Conference (BMVC)*, 2020.

[4] Steven A. Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L. Hammer, and Michael A. Riegler. The endotect 2020 challenge: Evaluation and comparison of classification, segmentation and inference time for endoscopy. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *ICPR International Workshops and Challenges*, pages 263–274, 2021. ISBN 978-3-030-68793-9.

[5] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Neural Information Processing Systems (NeurIPS)*, pages 5580–5590, 2017.

[6] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[7] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

[8] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[10] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1147.

[11] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2020.

[12] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *arXiv preprint arXiv:2012.15029*, 2020.

[13] B. N. Patro, M. Lunayach, and V. P. Namboodiri. Uncertainty class activation map (u-cam) using gradient certainty method. *IEEE Transactions on Image Processing*, 30: 1910–1924, 2021. doi: 10.1109/TIP.2020.3046916.

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[15] Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[16] 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. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 24–25, 2020.

[17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

[18] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[19] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32:10967–10978, 2019.

[20] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.