# Towards a Quantitative Analysis of Class Activation Mapping for Deep Learning-based Computer-Aided Diagnosis

Hanul Kang[a], Ho-min Park[a,b], Yuju Ahn[a], Arnout Van Messem[a,c], and Wesley De Neve[a,b]

[a]Center for Biotech Data Science, Department of Environmental Technology, Food Technology and Molecular Biotechnology, Ghent University Global Campus, Incheon, Korea
[b]IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium
[c]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[*Original version] The authors confirm that this work never has been published nor presented before.

## 1. DESCRIPTION OF PURPOSE

Deep learning using convolutional neural networks (CNNs)[1] has been successfully applied for the purpose of image classification and segmentation. In particular, within the field of healthcare, CNNs are currently frequently used to diagnose diseases,[2–7] with the correct treatment of a disease starting with an accurate diagnosis. Often, such a diagnosis relies on medical imaging such as X-ray, MRI, or PET scans. With the help of CNNs, these medical images can be automatically analyzed, aiding medical experts in decision-making by significantly reducing the time and effort needed for the detection and classification of abnormalities. However, CNNs are so-called black-box models, providing limited insight into their internal working. Nevertheless, the introduction of Class Activation Mapping (CAM)[8] made it possible to better understand how CNNs process images, and ever since, numerous studies have leveraged CAM to verify how well CNNs are doing in analyzing medical images. Table 1 provides an overview of a number of representative research efforts. However, most of these research efforts only use CAM for a qualitative model assessment, through the creation of heatmaps. In this paper, we discuss an approach that uses CAM for a quantitative model analysis, measuring how well CAM can be used to segment a brain tumor in MRI images, given that the type of brain tumor has been correctly classified by a CNN-based deep learning model.

## 2. MATERIALS AND METHODS

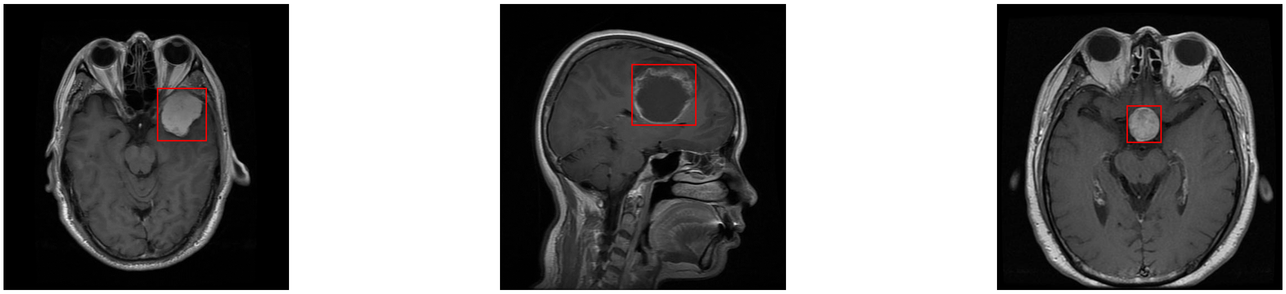
### 2.1 Dataset description

We make use of the brain tumor dataset[9] that was acquired between 2005 and 2010 at two Chinese hospitals, namely the Nanfang Hospital in Guangzhou and the General Hospital of Tianjing Medical University. This dataset consists of 3,064 T1-weighted brain MRI images stemming from 233 patients. Each image has been given a multi-class brain tumor label and is accompanied by its corresponding segmentation mask (ground

| Authors | Use case | Major model | Visualization |
|---|---|---|---|
| Rajpurkar et al.[2] | Pneumonia detection using chest X-rays | DenseNet | CAM |
| Bien et al.[3] | Knee injury classification using MRI scans | AlexNet | CAM |
| Han et al.[4] | Skin disorder classification using clinical skin images | ResNet | Grad-CAM |
| Nguyen et al.[5] | Eye tumor segmentation using MRI scans | ResNet and U-Net | CAM |
| Kiani et al.[6] | Liver cancer detection using biological tissue images | DenseNet | CAM |
| Kim et al.[7] | Glaucoma classification using fundus images | Inception-v4 | Grad-CAM |

Table 1. CNN-based models for computer-aided diagnosis in medical images and corresponding visualization methods

Further author information: (Send correspondence to Ho-min Park)
Ho-min Park: E-mail: homin.park@ugent.be



(a) Meningioma tumor (b) Glioma tumor (c) Pituitary tumor

Figure 1. Example image for each type of brain tumor. A red box is used to delineate a tumor.

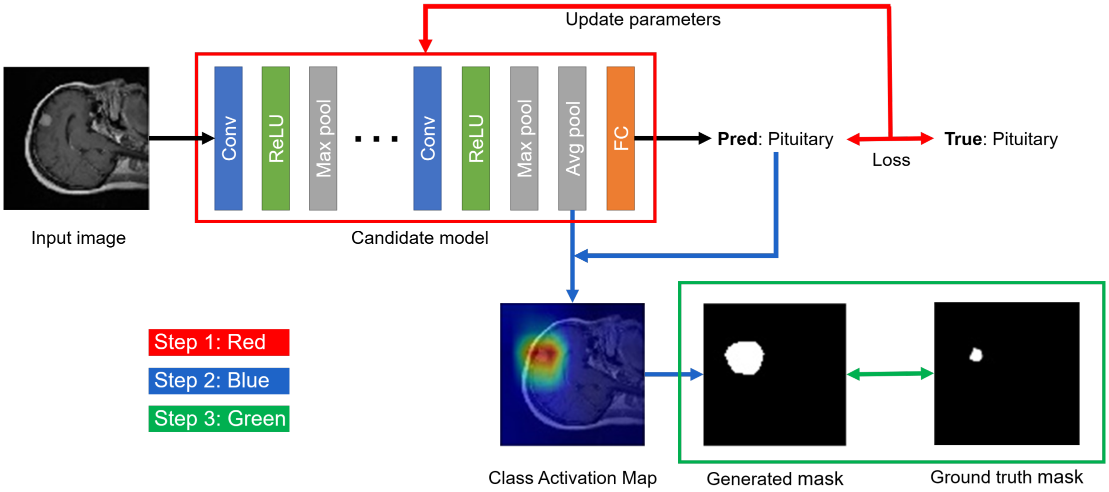


Figure 2. Overview of the proposed approach, leveraging a CNN that consists of convolutional layers (Conv), rectified linear unit layers (ReLU), pooling layers (average and maximum), and fully connected layers (FC).

thruth). A distinction is made between three types of tumors: meningioma (708 images), glioma (1,426 images), and pituitary (930 images). An example of each tumor type can be found in Figure 1. A meningioma tumor, which can be found in-between the skull and the brain, originates from the arachnoid cells that surround the brain. A glioma tumor, which can be found anywhere inside the skull, originates from the glioma cells that deliver the substances needed for the functioning of the neurons. A pituitary tumor is usually found near the center of the brain, occurring in the pituitary gland that governs the secretion of hormones. Furthermore, MRI images are, in general, divided over three anatomical planes: the sagittal, the axial, and the coronal plane. However, the adopted dataset does not contain any information about the anatomical plane used. Note that the images presented in Figure 1(a) and Figure 1(c) were created along the axial plane, while the image presented in Figure 1(b) was created along the sagittal plane.

## 2.2 Experimental approach

To perform a quantitative evaluation of the effectiveness of CAM in the field of medical imaging, we designed an experiment that consists of three different steps, as shown in Figure 2: (1) training a CNN-based candidate model, (2) generation of the class activation maps (heatmaps) and the segmentation masks, and (3) analysis of the effectiveness of segmentation.

The first step, training a CNN-based candidate model, is indicated in red in Figure 2. In this step, the task of finding the best classification model for the target dataset is addressed. When an image is given as an input to a candidate model, then the model predicts the type of tumor (label) present in the image. The difference between the predicted and the true label (loss) is calculated. This loss is then used for updating the parameters in the model in order to obtain a more accurate prediction, and thus a lower loss. We used three candidate

| Model | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| AlexNet | 0.902 (0.086) | 0.900 (0.091) | 0.896 (0.089) | 0.896 (0.089) |
| VGG-16 | 0.928 (0.089) | 0.926 (0.082) | 0.928 (0.072) | 0.924 (0.087) |
| VGG-19 | 0.950 (0.040) | 0.944 (0.048) | 0.946 (0.038) | 0.944 (0.043) |

Table 2. Mean (standard deviation) of the different classification performance metrics for the candidate models.

models: (1) AlexNet,[1] (2) VGG-16,[10] and (3) VGG-19.[10] An average pooling layer was added to AlexNet before the fully connected layer. For each of the VGG models, the $7 \times 7$ average pooling layer was changed into a $1 \times 1$ average pooling layer in order to be able to create a CAM in the next step. To mitigate the risk of overfitting, we applied transfer learning,[11] initializing the model parameters with values that have already been learned from other large datasets. To determine the effectiveness of each model on the given dataset, we used 5-fold cross validation. In other words, all images are used for testing at least once. Since the total number of images (3,064) is not divisible by 5, the dataset was split into four subsets of 613 images, and one subset of 612 images. That way, five instances were constructed for each of the three candidate models. The average effectiveness of each candidate model can be found in Table 2. Since VGG-19 comes with the highest classification effectiveness, we finally choose this model for the generation of class activation maps and segmentation masks.

The second step, as indicated in blue in Figure 2, consists of generating the CAM and the segmentation mask for a particular input image. The CAM is created by combining the output of the average pooling layer and the predicted tumor type. Furthermore, to obtain a binary segmentation mask, we make use of the 95th percentile value in the CAM as a threshold. This value was chosen because it comes with the highest average Intersection over Union (IoU), compared to the mean, median, 75th percentile, and 90th percentile value:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{1}$$

In a next step, we binarize the CAM image based on this threshold. In other words, if the pixel value in the CAM is higher than the threshold, it is set to white (255), otherwise it is set to black (0).

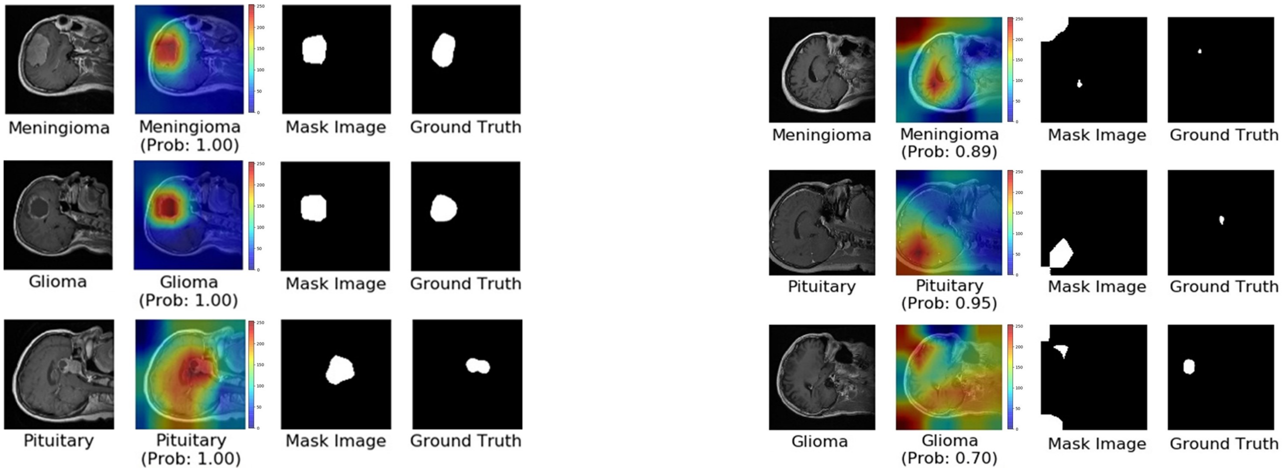
Finally, to analyze the effectiveness of segmentation (indicated in green in Figure 2), the overlap between the mask generated by CAM and the ground truth mask is investigated, using the IoU as the evaluation metric.

## 3. EXPERIMENTAL RESULTS

Table 2 summarizes the mean classification effectiveness obtained by the different candidate models. VGG-19, which performs best for every metric, was selected as the final classification model, and is thus used to generate the segmentation masks (in combination with CAM). A number of example images can be found in Figure 3, showing both classification and segmentation outcomes. Although VGG-19 is able to obtain a high classification effectiveness, the average IoU is low, only having a value of 0.153. Furthermore, only 23% of the meningioma tumors, 5% of the glioma tumors, and none of the pituitary tumors have an IoU larger than 0.5. Figure 3(b) shows a number of example images where the model failed to segment the tumor area, even though the tumor is clearly visible to the naked eye. Considering an IoU less than 0.01 as a badly segmented tumor, 42% of the tumors in the test set (257 cases out of 612) fall into this category, with the model not being able to correctly segment 85% of the pituitary tumors, 33% of the glioma tumors, and 2% of the meningioma tumors. Moreover, even if the model was able to correctly segment a tumor, we could observe a significant difference between the calculated segmentation mask and the ground truth mask. In particular, we could observe that the model was not able to precisely segment many small tumors (that is, tumors only about the size of a pea).

## 4. CONCLUSIONS AND FUTURE RESEARCH

Using CAM as a proxy for image segmentation needs to be done in a careful way, given that CAM learns from class labels and not from pixel-wise labels (as available in a binary segmentation mask). Nonetheless, by using CAM as a proxy for image segmentation, we found that (1) a predictive model does not always focus on the correct areas in an image to perform classification and that (2) a high classification effectiveness can still be paired



(a) Good segmentation

(b) Bad segmentation

Figure 3. Visual comparison between generated segmentation masks and ground truth masks.

with a low segmentation effectiveness (given the observed classification effectiveness of 95% versus an average IoU of only 0.153). Therefore, healthcare practitioners applying deep learning-based classification models to medical images need to be careful when interpreting the obtained model results.

In the near future, additional experiments will be performed for each step shown in Figure 2:

**Candidate models** – In this study, only two types of classification models (AlexNet and VGG) were used. These types of classification models were released in 2012 and 2014, respectively. In future research, we plan to investigate the use of ResNet,[12] DenseNet,[13] and more recently released models.

**Generation of CAMs and segmentation masks** – As an alternative to CAM, several follow-up techniques could be considered, such as Grad-CAM[14] and Guided Attention Inference Networks.[15] Also, binarization could be implemented using a number of alternative approaches, like Otsu's method,[16] which uses the global distribution of image brightness, and locally adaptive thresholding techniques.[17]

**Evaluation metrics** – In this study, the segmentation masks generated by CAM were evaluated by making use of IoU. However, as shown in Figure 3(a), even if a generated segmentation mask includes the ground truth region, the IoU is still low if, for example, the ground truth region is small compared to the predicted segmentation mask. Hence, a new metric is needed that is able to take into account such conditions.

## REFERENCES

[1] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," in [*Advances in Neural Information Processing Systems 25*], Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., eds., Curran Associates, Inc. (2012).

[2] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225* (2017).

[3] Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B. N., Yeom, K. W., Shpanskaya, K., et al., "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet," *PLoS Medicine* **15**(11), e1002699 (2018).

[4] Han, S., Kim, M., Lim, W., Park, G., Park, I., and Chang, S., "Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm," *Journal of Investigative Dermatology* **138**(7), 1529–1538 (2018).

[5] Nguyen, H.-G., Pica, A., Hrbacek, J., Weber, D. C., La Rosa, F., Schalenbourg, A., Sznitman, R., and Cuadra, M. B., "A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps," in [*International Conference on Medical Imaging with Deep Learning*], 370–379 (2019).

[6] Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C. P., Ball, R. L., Montine, T. J., et al., "Impact of a deep learning assistant on the histopathologic classification of liver cancer," *npj Digital Medicine* **3**(1), 1–8 (2020).

[7] Kim, M., Han, J., Hyun, S., Janssens, O., Van Hoecke, S., Kee, C., and De Neve, W., "Medinoid: Computer-Aided Diagnosis and Localization of Glaucoma Using Deep Learning," *Applied Sciences* **9**(15), 3064 (2019).

[8] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., "Learning Deep Features for Discriminative Localization," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 2921–2929, IEEE (2016).

[9] Cheng, J., *Brain tumor dataset* (2017 (accessed July 28th, 2020)). https://doi.org/10.6084/m9.figshare.1512427.v5.

[10] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556* (2014).

[11] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C., "A Survey on Deep Transfer Learning ," in [*International Conference on Artificial Neural Networks*], 270–279, Springer (2018).

[12] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition ," in [*Proceedings of IEEE International Conference on Computer Vision*], 770–778 (2016).

[13] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q., "Densely Connected Convolutional Networks," in [*Proceedings of IEEE International Conference on Computer Vision*], 4700–4708 (2017).

[14] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in [*Proceedings of IEEE International Conference on Computer Vision*], 618–626 (2017).

[15] Li, K., Wu, Z., Peng, K.-C., Ernst, J., and Fu, Y., "Tell Me Where to Look: Guided Attention Inference Network," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 9215–9223 (2018).

[16] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1), 62–66 (1979).

[17] Li, S. Z. and Jain, A., eds., [*Local Adaptive Thresholding*], 939–939, Springer US, Boston, MA (2009).