UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Weakly-supervised tumor segmentation in computed tomography scans

Author:
Iryna ZAKHARCHENKO

Supervisor:
Dmytro FISHMAN

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2023

Declaration of Authorship

I, Iryna ZAKHARCHENKO, declare that this thesis titled, "Weakly-supervised tumor segmentation in computed tomography scans" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"If you set your goals ridiculously high and it's a failure, you will fail above everyone else's success. "

James Cameron

"Life is what happens when you're busy making other plans. "

John Lennon

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Weakly-supervised tumor segmentation in computed tomography scans

by Iryna ZAKHARCHENKO

Abstract

This research focuses on the topic of weakly-supervised tumor segmentation. The proposed pipeline involves the usage of a classification model to make predictions regarding the presence of a tumor in an image. Subsequently, the CAM (Class Activation Mapping) approach is employed to identify the most relevant regions within the image as determined by the model. The underlying concept is that the model will learn to identify tumor regions, resulting in higher activations in those areas. The advantage of the weakly supervised approach is its ability to learn from a smaller dataset, requiring only image-level labels in our case. By implementing the proposed pipeline, specifically using the Score-CAM technique.

Acknowledgements

I would like to express my sincere gratitude to my mentor, Dmytro Fishman, for his solid support and invaluable suggestions throughout this research. I am deeply thankful to the Applied Sciences Faculty at the Ukrainian Catholic University, particularly Oleksii Molchanovsky, for their guidance and assistance. I am also grateful to the Armed Forces of Ukraine for their brave service.

Lastly, I would like to extend my heartfelt thanks to my family, friends and fiancé for their constant support and understanding during the course of this research.

Contents

List of Figures

List of Tables

List of Abbreviations

CNN	**C**onvolutional **N**eural **N**etwork
ResNet	**R**esidual **N**etwork
IoU	**I**ntersection over **U**nion
CAM	**C**lass **A**ctivation **M**apping
CT	**C**omputed **T**omography
ReLU	**R**ectified **L**inear **U**nit
LSTM	**L**ong **S**hort-**T**erm **M**emory

Dedicated to my father

Chapter 1

Introduction

In the medical field, the ability of doctors to accurately identify diseases can sometimes be compromised due to long working hours and stress. Important details in medical images, such as computed tomography (CT) scans, may be accidentally overlooked, leading to potentially harmful consequences. The increasing workload on healthcare professionals further inflames this issue. In light of these challenges, our solution aims to assist doctors in detecting and localizing tumors and cysts in CT images. Early detection of even the smallest indications of cancer in organs like the lungs, kidneys, or breast is crucial for timely intervention and improved patient outcomes.

Our proposed solution involves the development of a model capable of detecting and classifying cancer using a weakly-supervised approach. This approach requires lower-quality data, specifically image-level labels indicating the presence of a tumor. While segmentation requires the tumor's mask, which is more challenging to obtain, we demonstrate that the weakly-supervised approach can effectively train a classification model for tumor segmentation without the need for pixel-level annotations.

Furthermore, we see the potential to collect additional information about a patient's symptoms or familial illnesses, enabling better prediction and risk assessment for individual cases.

1.1 Hypothesis

We put four of the following hypotheses for our research:

1. Weakly-supervised approaches have the potential for tumor segmentation and can achieve promising results.

2. Score-CAM, a specific CAM approach, can outperform Grad-CAM approaches in terms of performance.

3. Given the limited amount of available data, employing shallower network architectures holds the potential for yielding improved results in our task.

4. Supervised approaches will outperform weakly-supervised approaches in terms of segmentation accuracy, as they directly learn the correct segmentation.

1.2 Thesis structure

The structure of our thesis is outlined as follows:

Background This chapter provides an overview of the medical and technical background, including computed tomography scans, tumor characteristics, computer vision tasks, class activation mapping techniques, relevant metrics, and previous research conducted in the field.

Materials and methods This chapter describes the datasets utilized in our research, including Deeplesion, Kits-19 and Kits-21. We outline the weakly-supervised approach adopted in our experiments.

Experiments and results This chapter presents the experiments conducted using various Class Activation Mapping approaches and reports the corresponding results.

Conclusion In this concluding chapter, we summarize the findings and results of our research. We also outline potential future steps that were not explored in this work but hold promise for improving the results.

Chapter 2

Background information

This chapter aims to provide a comprehensive overview of the medical and technical backgrounds relevant to the research conducted. The medical background section will focus on essential aspects related to CT and tumors. By examining CT imaging techniques and their significance in tumor detection, a foundation will be established for the subsequent research.

In the technical background section, the emphasis will be on computer vision tasks and their relevance to the current study. This will involve a discussion of the various tasks within computer vision, such as image classification and segmentation, which are crucial to tumor analysis and detection. Furthermore, a review of related work in the field will be presented to highlight existing methodologies and approaches.

Another key aspect addressed in the technical background section will be the introduction and explanation of CAM (Class Activation Mapping) methods. These techniques play a pivotal role in the research methodology and are instrumental in identifying the most crucial regions within an image as determined by the model. A thorough overview of CAM methods will be provided to ensure a clear understanding of their significance and application within the context of the study.

2.1 Medical background

This chapter serves to provide an overview of the pivotal medical aspects that are applicable to our research. Gaining an understanding of the medical background is crucial for the data collection process in real-world scenarios. Our discussion will primarily revolve around two significant topics: computed tomography and tumors.

The section dedicated to CT will dive into the principles and techniques employed in this imaging modality. In addition, we will direct our attention towards tumors, a focal point of our research. This will involve an exploration of the characteristics and types of tumors within the medical context. By delving into the diverse aspects of tumors, we aim to establish a solid foundation for their detection and segmentation.

2.1.1 Computed tomography

Computed tomography (CT) is an essential medical imaging technique employed in radiology for diagnostic purposes. It involves directing a narrow beam of X-rays towards a patient, which swiftly rotates around their body. These X-rays generate signals that are subsequently processed by a computer to produce "slices" of the body. By combining all the collected slices digitally, a three-dimensional (3D) image of the patient is formed.

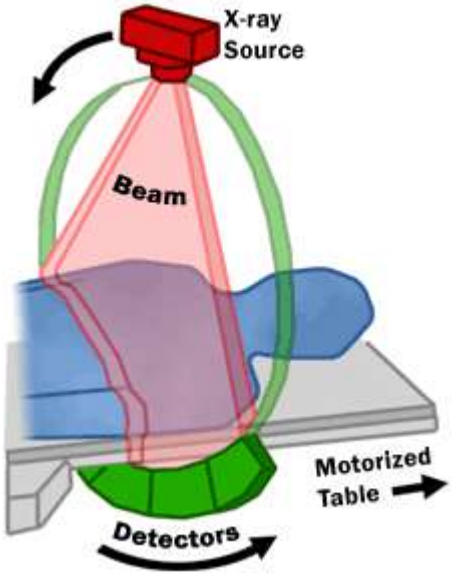

FIGURE 2.1: The process of data collection through a CT scan. *CT scan* 2023

The thickness of the tissue represented in each image slice can vary, depending on the specific CT machine utilized, typically ranging from 1 to 10 millimeters.

The data obtained from this procedure is in the form of voxels, which represent three-dimensional volumes of the human body. However, for our approach, we focus on working with images, and therefore we need to extract horizontal slices from the voxel data. Conceptually, we can envision each slice as a single circular path of the X-ray source around the body. Consequently, a single slice may depict, for example, both kidneys and a tumor in the same image. By leveraging these horizontal slices, we can effectively analyze and interpret the medical images as part of our research approach.

2.1.2 Tumor

A tumor refers to a collection or cluster of abnormal cells that manifest within the body. It is important to note that not all tumors are cancerous; some are classified as noncancerous or benign. *Kidney Cancer* 2023

Tumors typically emerge when cells undergo excessive division and growth within the body. Ordinarily, the body maintains regulation over the growth and division of cells. Aging cells are naturally replaced by new ones, ensuring the proper functioning of bodily processes. Moreover, damaged cells are replaced by healthy ones. However, when this regulatory process becomes disrupted or corrupted, healthy cells may undergo changes and proliferate uncontrollably, ultimately forming a mass commonly referred to as a renal cortical tumor.

There are indeed several types of kidney cancer. The most common types include:

- Renal cell carcinoma. One of the most common kidney cancer in adults, near 85% of diagnoses.
- Urothelial carcinoma. It accounts for 5% to 10% of the kidney cancers diagnosed in adults.

- Sarcoma. Sarcoma of the kidney is rare. This type of cancer develops in the soft tissue of the kidney.
- Wilms tumor. Wilms tumor is most common in children. Wilms tumors make up about 1% of kidney cancers.
- Lymphoma. Lymphoma can enlarge both kidneys and is associated with enlarged lymph nodes. *Kidney Cancer* 2023

Cysts, conversely, are mostly benign in nature. They do not exhibit any harmful effects on kidney function and do not result in organ enlargement. *Simple Kidney Cysts* 2019

Within the utilized dataset, labels were provided for both tumors and cysts. While our primary objective does not revolve around differentiating between these two conditions, our aim is to detect the presence of either of these diseases within an image sample.

2.2 Technical background

Our main goal in this chapter is to examine the technical details of our project. We will concentrate on diverse computer vision tasks, the use of class activation mapping (CAM) methods, the assessment of metrics, and a review of related research in the area.

2.2.1 Computer vision tasks

Computer vision, a multidisciplinary field at the intersection of computer science and image processing, plays a pivotal role in the medical domain. It incorporates a range of algorithms, methodologies, and techniques designed to extract meaningful information and insights from medical images, ultimately aiding in diagnosis and disease monitoring.

Within the field of medicine, computer vision is used for advanced image analysis and pattern recognition algorithms to interpret medical images acquired from diverse modalities such as X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and histopathology slides. By applying sophisticated algorithms, computer vision enables the automated detection, segmentation, and classification of anatomical structures, lesions, tumors, and other pathological abnormalities.

The application of computer vision in medicine holds substantial potential for enhancing clinical decision-making and patient care. It can assist radiologists, pathologists, and other healthcare professionals by providing quantitative measurements, aiding in the identification of subtle anomalies, and classification of diseases. Moreover, computer vision techniques can contribute to the efficient analysis of large-scale medical image datasets, accelerating research.

This research is primarily centered around two fundamental computer vision tasks: classification and segmentation.

Image classification

Image classification is a supervised approach in which a model is trained to categorize objects within an image. The model receives an input image and is provided

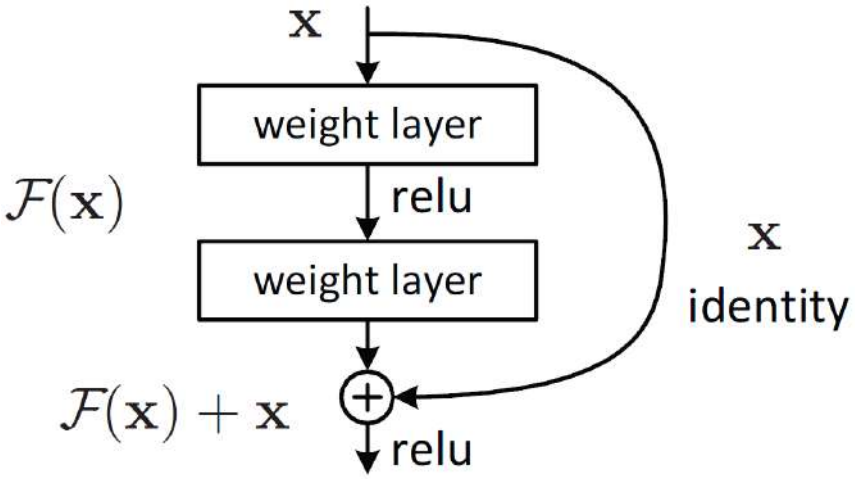

FIGURE 2.2: Residual learning: a building block

with specific, predefined classes as labels. The goal is to enable the model to accurately assign the correct label to new, unseen images.

To train a robust classification model, it is essential to have a diverse and representative dataset that includes examples from all the classes we aim to distinguish. A greater variety of images for each class allows the model to learn generalized patterns rather than focusing solely on specific features present in individual examples.

In our research, the focus lies specifically on the classification of images containing kidney tumors. We have narrowed our investigation to kidney tumor. However, it is important to note that the approach we develop and the insights gained from our research may potentially be applied to other organs as well. The underlying principles and techniques can be adapted and extended to address classification challenges in different medical contexts, expanding the scope of our findings and their potential impact.

Residual Network architecture The researchers in this study observed that increasing the size of the models led to improved performance in their classification task. However, they encountered a problem known as the vanishing gradient, wherein the gradients propagated through the deep layers of the model became increasingly small, making learning less effective.

To address this issue, the researchers proposed a solution: incorporating skip connections within the model architecture. That allow information from earlier layers to bypass subsequent layers and directly influence the final prediction. This process is visually represented in Figure 2.2.

The core building block of the Residual Network (ResNet) architecture He et al., 2015 is the residual block, which consists of two convolutional layers and a skip connection. The input to the block is denoted as x, and the convolutions produce a transformed output denoted as $F(x)$. Simultaneously, the skip connection directly passes the original input x to the next layer. By introducing these skip connections, the model can effectively address the vanishing gradient problem and accommodate deeper network structures, thereby enabling the use of larger and more powerful models.

ResNet-50, a variant of the ResNet, exhibits a similar structure to ResNet-34 but incorporates an additional bottleneck design.

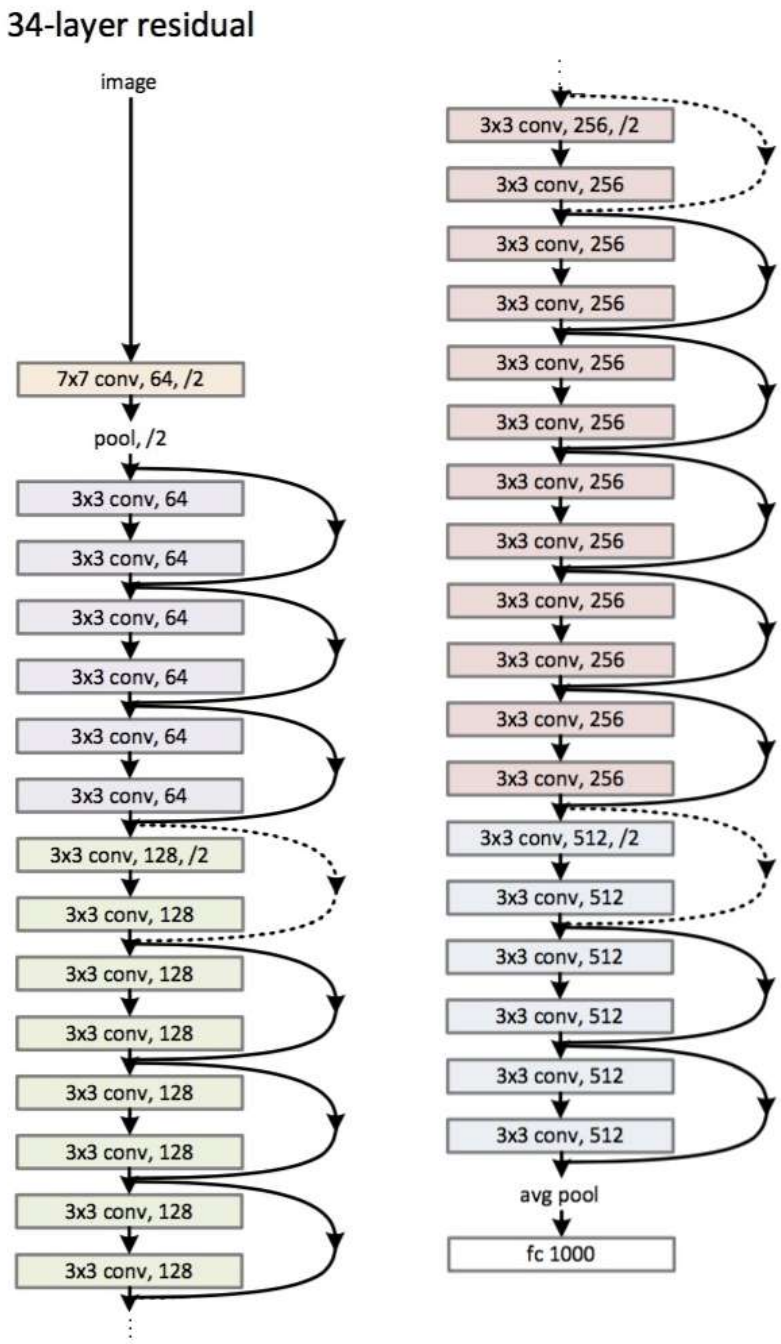

FIGURE 2.3: ResNet arhitecture

The inclusion of the bottleneck layer allows ResNet-50 to strike a balance between model depth, expressiveness, and computational efficiency. It enables the model to capture both local and global features effectively, facilitating accurate and robust predictions while optimizing computational resources.

In both ResNet-50 and ResNet-101 architectures, the authors incorporate 1x1 convolutions for reducing the number of features. Other great characteristic of 1x1 convolutions is that they are able to transform the number of channels in the input tensor. It can be used to increase or decrease the number of channels, thereby controlling the complexity and capacity of the network. Additionally, it helps to introduce non-linearity through the activation function applied after the convolution operation.

Semantic segmentation

Semantic segmentation is a fundamental task in computer vision that involves assigning a specific class label to each pixel in an image. In the context of our research, these classes can represent various elements such as tumors, cysts, kidneys, lungs, and more. When multiple elements overlap within a single image, our approach does not differentiate between them; instead, the model generates a merged mask that encompasses all detected instances.

The outcome of semantic segmentation is a detailed mask that designates the specific pixels belonging to different classes. This not only allows us to determine the presence of a tumor in an image but also enables us to predict its precise mask. Using this mask, we can perform additional calculations, such as measuring the size and intensity of the tumor pixels or counting the number of tumors present, provided they are not overlapping. Alternatively, the prediction may manifest as a single point at the center of the tumor, effectively highlighting the affected region.

By employing semantic segmentation techniques, we can achieve a more granular understanding of medical images, enabling us to extract valuable information for diagnosis, treatment planning, and further analysis. This approach holds great potential for advancing the field of medical imaging and improving patient care by facilitating the accurate localization and characterization of various abnormalities within the images.

Encoder-Decoder Architecture The architecture used in this study employs an Encoder-Decoder framework, consisting of two primary components:

- Encoder: This component, represented by the left portion of Figure 2.4, is responsible for extracting essential information from the input image. By utilizing various layers and operations, the encoder produces a high-dimensional feature vector that captures the salient features of the image.
- Decoder: The decoder, depicted by the right portion of Figure 2.4, performs the reverse process by decoding the information from the bottleneck representation and generating a semantic map. This map provides a detailed understanding of the image, highlighting different regions and their respective classes.

The high-dimensional feature vector obtained from the encoder stage serves as a valuable representation of the input image. It can be further utilized for tasks such as image reconstruction in autoencoders or generating image segmentation masks. By leveraging this architecture, we aim to enhance the understanding and analysis

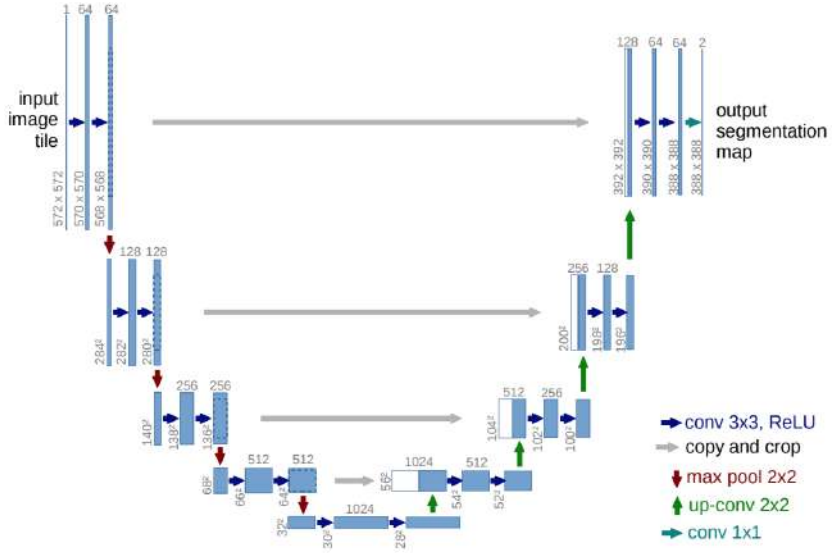

FIGURE 2.4: U-Net arhitecture

of images, enabling accurate and effective segmentation and mapping of various elements within the image.

A prominent instance of the Encoder-Decoder architecture is the U-Net model Ronneberger, Fischer, and Brox, 2015, as illustrated in Figure 2.4. The U-Net architecture incorporates skip connections, which effectively address the segmentation of smaller objects and yield outputs with improved smoothness. Furthermore, the model demonstrates a relatively low number of parameters, rendering it suitable for training on limited datasets, particularly when combined with appropriate augmentation techniques. Notably, the U-Net architecture has garnered significant success and widespread adoption in the field of biomedical imaging, specifically for segmentation tasks Ronneberger, Fischer, and Brox, 2015; Liu et al., 2022; Hollo, 2019.

2.2.2 Class Activation Mapping

Class Activation Mapping (CAM) Zhou et al., 2015 is a valuable technique applied in convolutional neural networks (CNNs) to highlight regions of interest related to specific classes. This approach offers insights into the decision-making process of the model and serves as a powerful tool for model improvement, enabling a visual understanding of the areas on which the model focuses. Even in cases where the model is not explicitly trained in a supervised manner to localize specific classes, it tends to capture patterns from images containing those classes. Consequently, the model develops an understanding of where the objects of interest are likely to exist within an image. These patterns can manifest as concrete features shared among different class representations. For instance, when dealing with the "dog" class, the model attempts to identify characteristic elements such as ears, eyes, nose, and tails. Dogs may vary in terms of breeds, colors, and sizes, further highlighting the need for the model to differentiate between different classes, such as "dog" and "cat," based on distinct features like muzzle shape or ear appearance.

CAM techniques aim to stress the regions where the model identifies specific cues related to the target class, making it a valuable tool for visual interpretation and analysis. This is precisely why CAM is of significant interest in the field.

However, it is important to note that the CAM method has a particular limitation: the model architecture must incorporate global pooling as the final layer and should not contain fully connected layers prior to that, similar to the VGG architecture.

Let us denote f as the CNN model with global pooling. Given an input X, the model produces a prediction Y, representing a probability distribution over all classes C, with Y_c indicating the confidence score for class c. Additionally, we denote A_l as the activation of layer l. In the case of a convolutional layer, the activation is denoted as A_l^k, where k represents the channel. The weight between the k-th neuron of layers l and $l+1$ is denoted as $w_{l,l+1}$.

Definition 1 (Class Activation Mapping)
Class activation mapping is a technique applicable to models featuring global pooling. In this context, a model with global pooling at layer l uses the output of the preceding layer, $l-1$, as input and passes the pooled activation to a fully connected layer, $l+1$. Specifically, for a given class c, L_{CAM}^c can be defined as follows:

$$L_{CAM}^c = ReLU(\sum_k \alpha_k^c A_{l-1}^k)$$

, where $\alpha_k^c = w_{l,l+1}^c[k]$

The ReLU is used for calculating only positive correlations with class c prediction. The driving force behind CAM is the fact that each activation map Alk carries some unique dimensional information about input X and the weight of each channel is linear combination of fully connected layer and global pooling. As we said previously this approach can not be used to all architectures.

Grad-CAM is an approach that addresses the limitation of CAM. This method was proposed by Selvaraju et al. in their paper "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization" Selvaraju et al., 2019. The key idea behind Grad-CAM is to capture the importance of each feature map in the network by computing the gradients of the class confidence score with respect to the feature maps. These gradients represent the sensitivity of the class prediction to changes in the feature maps.

Grad-CAM redefines α_k^c as a gradient of class confidence Y^c . With such approach, we can use models that do not have global pooling.

Definition 2 (Grad-CAM)
For Grad-CAM, consider a convolutional layer $l$in a model f, and having class of interest c:

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A_{l-1}^k)$$

, where $\alpha_k^c = GP(\frac{\partial Y^c}{\partial A_l^k})$

Visual representation of Grad-CAM pipeline is visualized on Figure 2.5.

The Grad-CAM++ technique by Chattopadhay et al., 2018 is a variant of Grad-CAM that introduces a redefined combination of gradients. It shares similarities

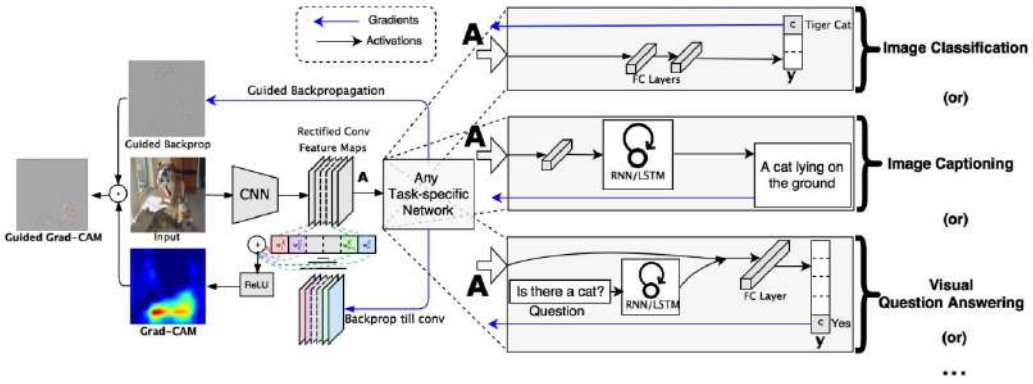

FIGURE 2.5: Grad-CAM pipeline

with Grad-CAM but differs in the specific formulation used to combine gradients for generating the sensitivity maps.

Smooth Grad-CAM++ by Omeiza et al., 2019 is an extension of Grad-CAM++ that aims to enhance the sharpness of gradient-based sensitivity maps. It achieves this by introducing random sampling in the neighborhood of an input x and subsequently averaging the resulting sensitivity maps, thereby refining the overall quality and clarity of the generated maps.

On the other hand, ScoreCAM by Wang et al., 2020 is a gradient-free method that operates by producing heatmaps corresponding to different regions of an image. It then uses only those specific areas during the model prediction process. This approach allows researchers to identify the most crucial heatmaps that contribute significantly to achieving optimal predictions. The measure of importance is captured through the concept of Increase of Confidence, which is incorporated within the ScoreCAM methodology.

Definition 3 – Increase of Confidence

Given a general function $Y = f(X)$ that takes an input vector $X = [x_0, x_1, ..., x_n]$ and outputs a scalar Y . For a known baseline input X_b, the contribution c_i of x_i, $(i \in [0, n-1])$ towards Y is the change of the output by replacing the i-th entry in X_b with x_i . Formally,

$$c_i = f(X_b \circ H_i) - f(X_b)$$

where H_i is a vector with the same shape of X_b but for each entry h_j in H_i ,

$h_j = \mathbb{I}[i = j]$ and $\circ$ denotes Hadamard Product.

Definition 4 - CIC score

Given a CNN model $Y = f(X)$ that takes an input X and outputs a scalar Y . We pick an internal convolutional layer l in f and the corresponding activation as A. Denote the k-th channel of A_l by A_l^k. For a known baseline input X_b, the contribution A_l^k towards Y is defined as

$$C(A_l^k) = f(X \circ H_l^k) - f(X_b)$$

where $H_l^k = s(Up(A_l^k))$,

$Up(\cdot)$ denotes the operation that upsamples A_l^k into the input size and $s(\cdot)$ is a normalization function that maps each element in the input matrix into $[0, 1]$.

Definition 5 (Score-CAM)

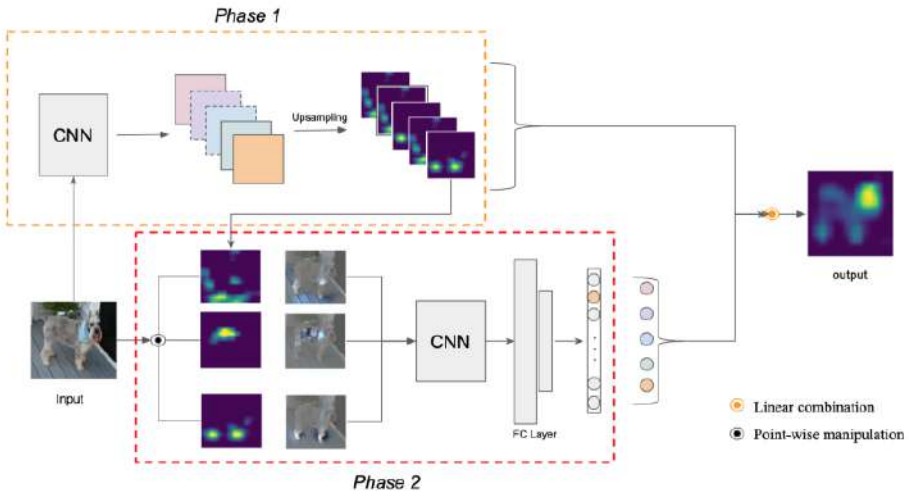

FIGURE 2.6: Score-CAM pipeline

Consider a convolutional layer l in a model f, given a class of interest c, Score-CAM $L^c_{Score-CAM}$ can be defined as

$$L^c_{Score-CAM} = ReLU(\sum_k \alpha^c_k A^k_l)$$

where $\alpha^c_k = C(A^k_l)$,
where $C(\cdot)$ denotes the CIC score for activation map A^k_l.

It is evident that in the case of ScoreCAM, the importance of the class activation maps is determined based on the class score rather than relying on the gradient information.

The proposed pipeline is depicted in Figure 2.6, illustrating the key steps of the approach:

- In the Phase 1, the input image undergoes processing by a CNN, resulting in the generation of feature maps from the final convolutional layer.
- In the Phase 2, the class score is calculated for each specific class by utilizing the heatmap as a mask. Notably, the importance scores are computed based on the class score rather than gradients.
- These importance scores are then employed as weights to perform a weighted summation of the activation maps obtained from the last convolutional layer. Each activation map is multiplied by its corresponding importance weight and subsequently added together.
- To focus on the positive contributions towards class prediction, ReLU activation is applied to the resulting weighted summation. This ensures that only the positive aspects of the feature maps are considered.

2.2.3 Metrics

In this section, we will provide an overview of the metrics that are commonly employed for evaluating both classification and segmentation tasks. These metrics play

a vital role as they enable quantitative comparison and assessment of different approaches.

Classification metrics

In the following sections, we will provide detailed descriptions of the main evaluation metrics used in our study: precision, recall, and $F1$-score. These metrics play a crucial role in assessing the performance of our models and understanding their strengths and weaknesses.

Precision and Recall Precision and recall are indeed important metrics for evaluating classification models. They provide valuable insights into the model's performance.

Precision is defined as the ratio of true positive predictions to the total number of positive predictions made by the model. Mathematically, precision is calculated as:

$$Precision = TP/(TP + FP)$$

where TP represents true positive predictions (correct prediction of positive class), FP represent false positive predictions (incorrectly predicted positive instances).

Recall, also known as sensitivity or true positive rate, is defined as the ratio of true positive predictions to the total number of actual positive instances in the dataset. It measures the proportion of correctly predicted positive instances out of all actual positive instances. Mathematically, recall is calculated as:

$$Recall = TP/(TP + FN)$$

where FN represents the number of false negative predictions (incorrectly predicted negative instances).

Precision focuses on the accuracy of positive predictions, indicating how precise the model is in correctly identifying positive instances. A high precision value indicates that the model has a low rate of false positives. Recall, on the other hand, focuses on the completeness of positive predictions, measuring the model's ability to capture all positive instances. A high recall value indicates that the model has a low rate of false negatives.

Both precision and recall are crucial metrics in evaluating the performance of a classification model, and they often exhibit a trade-off relationship.

Finding the right balance between precision and recall depends on the specific requirements and priorities of the classification task. For instance, in a medical diagnosis scenario, it may be more critical to have high recall to avoid missing potentially positive cases, even if it comes at the cost of lower precision. On the other hand, there exists cases when the precision is prioritized over recall.

F1-score The F1 score is a metric that combines both precision and recall into a single value, providing an overall assessment of the model's performance. It is calculated as the harmonic mean of precision and recall. The harmonic mean is used instead of a simple average because it gives more weight to lower values. This means that the F1 score is particularly sensitive to cases where either precision or recall is

low. It penalizes models that have a large disparity between precision and recall, encouraging a balanced performance. Mathematically, F1 score is calculated as:

$$F1score = 2 \cdot Precision \cdot Recall / (Precision + Recall)$$

The F1 score will be high only if both precision and recall are high, and it will be low if either precision or recall is low. It provides a balanced measure of the model's performance, considering both aspects of correct positive predictions and the ability to capture all positive instances.

Segmentation metrics

In the following paragraphs, we will provide an overview of segmentation metrics commonly used to quantify the performance of segmentation models, with a particular focus on the Intersection over Union (IoU) and Dice score.

Intersection over union The Intersection over Union (IoU) is a commonly used metric for evaluating segmentation models. It provides a measure of overlap between the predicted and ground truth regions. The IoU score is calculated by dividing the cardinality of the intersection of the predicted and ground truth regions by the sum of the cardinalities of these regions.

Mathematically, the IoU score is expressed as:

$$IOU = \frac{|X \cap Y|}{|X| + |Y|}$$

In this equation, X represents the predicted region and Y represents the ground truth region. The cardinality of a region is the number of elements it contains.

The IoU score ranges from 0 to 1, where a score of 1 indicates a perfect overlap between the predicted and ground truth regions, and a score of 0 indicates no overlap. By using the IoU metric, researchers and practitioners can quantitatively assess the performance of segmentation models by measuring the extent to which the predicted regions align with the ground truth regions.

Dice score The Sørensen–Dice coefficient (Dice score) is a widely used metric for evaluating segmentation models. It measures the similarity or overlap between the predicted and ground truth regions. The Dice score is calculated by taking twice the cardinality of the intersection of the predicted and ground truth regions and dividing it by the sum of the cardinalities of these regions.

Mathematically, the Dice score is expressed as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

$$DSC = \frac{2TP}{2TP + FN + FP}$$

In this equation, X represents the predicted region and Y represents the ground truth region.

The Dice score ranges from 0 to 1, with a score of 1 indicating a precise overlap between the predicted and ground truth regions, and a score of 0 indicating no overlap.

2.2.4 Related Work

In this section, we will explore various studies and research papers that focus on working with CT scans for tumor detection in different organs. While the specific focus of these studies may vary, they provide valuable insights and methodologies that can be adapted for our proposed approach.

In a research conducted by Kaspar Hollo, 2019, the author investigated weakly-supervised learning for artefact segmentation in brightfield microscopy images. A comparison was made between a supervised approach using U-Net and YOLOv5, and a weakly-supervised approach. The weakly-supervised pipeline involved training a ResNet50 architecture for binary classification of artefact presence, followed by using Score-CAM predictions as pseudo-labels on the pixel level and feeding them to the U-net model. The supervised approach achieved a pixel-wise F1-score of 0.92 and pixel-wise IOU of 0.86, while the weakly-supervised pipeline yielded a pixel-wise F1-score of 0.64 and pixel-wise IOU of 0.48.

In the project by Xinyang Feng et al., 2017 and colleagues, the focus was on the automatic segmentation of pulmonary disease in lung CT scans. Their proposed model generated voxel-level segmentation using image-level labels, allowing the model to identify areas of greater interest. This approach significantly reduced the amount of labeled data required for training, achieving segmentation accuracy comparable to benchmark fully supervised methods, especially for larger nodules.

Jinzheng Cai et al., 2017 proposed a work titled "Improving Deep Pancreas Segmentation in CT and MRI Images via Recurrent Neural Contextual Learning and Direct Loss Function." The researchers leveraged LSTM to incorporate information from not only the current image but also neighboring slices. The proposed method utilized a deep learning model, with its output serving as input to the LSTM subnetwork. The deep learning model was trained to optimize a segmentation loss function called Jaccard Loss. This approach outperformed other methods in the field of pancreas segmentation.

Sarah Ryan et al., 2020 worked on "Cluster Activation Mapping with Applications to Medical Imaging" in 2020. The research focused on dividing voxels into clusters and gaining insights into the differences between these clusters using Activation Mapping. The main characteristics identified for the clusters were the location of abnormality and absence of abnormality. However, this approach currently requires manual intervention to distinguish characteristics within each cluster.

Another relevant paper titled "Mixed-UNet: Refined Class Activation Mapping for Weakly-Supervised Semantic Segmentation with Multi-scale Inference" Liu et al., 2022, authored by Yang Liu, Ersi Zhang, Lulu Xu, and others, proposes an approach to improve segmentations in methods utilizing CAM. The researchers developed a new model called Mixed-UNet, which incorporates two parallel branches in the decoding phase. Experimental results demonstrated that the model outperformed other methods under the same supervision. This approach presents a potential avenue for further improving our pipeline.

In the paper "Weakly-supervised convolutional neural networks for renal tumor segmentation in abdominal CTA images" Yang et al., 2020, the authors propose a pipeline for segmenting renal tumors in abdominal computed tomography angiography (CTA) images. The pipeline takes an image as input, along with a bounding box around the tumor, and produces a tumor segmentation as output. The first step of the pipeline involves generating pseudo-labels using convolutional conditional random fields. This step helps in refining the initial segmentation by incorporating contextual information. Next, multiple CNN models are trained using a split of the

dataset into k subsets. Each model is trained on a different subset, resulting in k predictions for each image in the dataset. Finally, a final model is trained using the prepared masks obtained from the k predictions. This final model is then evaluated to assess the performance of the approach. The authors report achieving a Dice score of 0.826 using these methods. One notable difference between this approach and our proposed approach is that it requires a bounding box as input. This bounding box serves as a reference for the tumor region. This approach can be advantageous in scenarios where bounding box annotations are readily available and can aid in the labeling process.

Chapter 3

Materials and methods

We propose a pipeline for tumor segmentation using an weakly-supervised learning model. The pipeline involves training a ResNet model for tumor classification, where the model learns to classify images as either containing a tumor or not. Once the classification model is trained, we focus on the regions of the image that were most influential in the model's decision-making process. By examining these regions, we can determine the location of the tumor.

This pipeline offers a significant advantage as it allows us to process raw images and pinpoint the precise location of the tumor, without requiring explicit segmentation labels during training.

In our task, the input image corresponds to a slice from a CT scan. Further details on the selection of these slices will be discussed in later sections. The model's objective is to classify each image slice as either tumor-positive or tumor-negative. Following the classification, the CAM method is applied to highlight the areas of interest. By setting an appropriate threshold, we can perform the tumor segmentation.

A visual representation of our pipeline is presented in Figure 3.1.

3.1 Datasets

When selecting the primary research area of kidney tumor segmentation, our first step was to search for a suitable dataset that met our requirements. We needed a dataset that provided tumor segmentation or, at the very least, bounding boxes around the tumors. Additionally, we were interested in acquiring kidney images without tumors for the purpose of training a classification model.

One dataset we came across was DeepLesion, which consisted of 10,594 CT scans encompassing various organs. Although this dataset contained a substantial number of kidney tumors (495), it only provided the central image with a tumor and a corresponding bounding box. Consequently, we determined that the Kits-21 dataset would better suit our needs for kidney tumor segmentation, as it offered more comprehensive segmentation annotations.

3.1.1 Kits-19 and Kits-21

The Kits-19 and Kits-21 datasets Heller et al., 2020 are among the most widely used datasets for kidney tumor segmentation. The Kits-21 dataset, in particular, provides annotations for three labels: kidney, tumor, and cyst. In our research, we combine the tumor and cyst labels into a single label. While it is possible to separate them in the future, doing so may result in lower segmentation scores. This dataset consists of 100 CT scans from different patients. Both datasets share an identical set of voxels, ensuring consistency in the volumetric representation of the data. It is important to

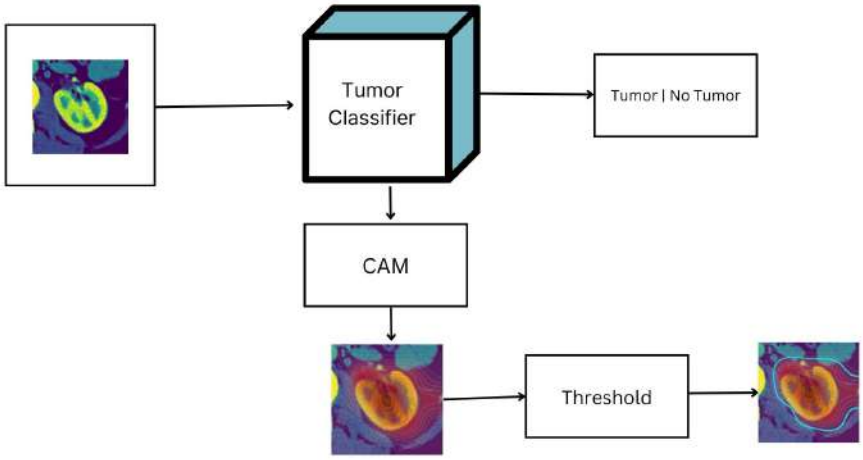

FIGURE 3.1: The proposed pipeline

mention that the Kits-21 dataset incorporates an update, specifically the inclusion of a cyst class, which distinguishes it from the Kits-19.

Although the dataset provides voxel data, we need to convert it into image format for further processing. To simplify the task for the model and enhance its performance, we split each image into left and right parts. By doing so, the model focuses on one kidney at a time. Additionally, we position the kidney in the center of a 150 x 150 image. This approach can be replaced by another model that precisely locates each kidney. However, using a smaller square ensures that the model's attention is primarily directed towards the kidney, even in cases where the kidney may have another affected organ nearby. The visual representation of the split is on Figure 3.2. By employing this approach, we are able to obtain a dataset that consists of 26,409 samples without tumor and 10,870 samples with tumor, both are visualized on Figure 3.3. In order to provide a comprehensive understanding of the data, we have prepared a graph 3.4 that illustrates the percentages of tumor coverage within the image. This graph serves as a statistical count, while for a visual representation we compose a Figure 3.5, both of the visuals are offering valuable insights into the extent of tumor presence in the dataset.

This approach allowed us to isolate and concentrate on individual kidney instances within the dataset. The resulting smaller images maintained the same resolution as the original images. This standardized size ensured consistency throughout the dataset, enabling efficient model training and evaluation processes.

All the data samples in the dataset exhibit a similar appearance. The dataset comprises CT scans collected from two hospitals over the period of 2010 to 2020, but there is no explicit indication of the specific hospital or temporal splitting within the dataset. Each patient in the dataset is represented by a unique number, and it is unknown whether there are multiple scans available for each individual. Visually, the scans appear distinct from one another, suggesting potential variations in imaging characteristics across the dataset.

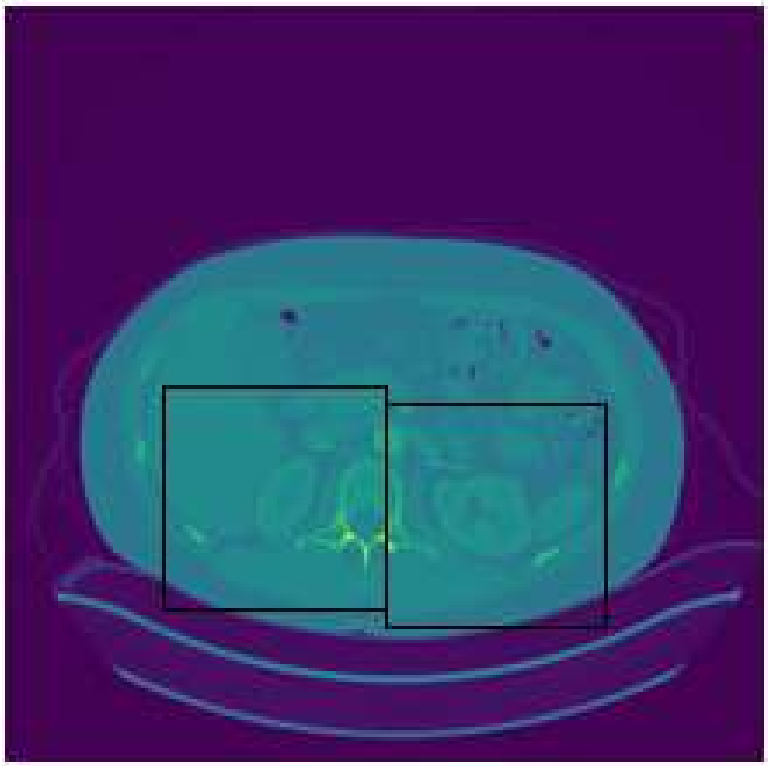

FIGURE 3.2: The example from the dataset with the two smaller cuts indicated by the black rectangles.

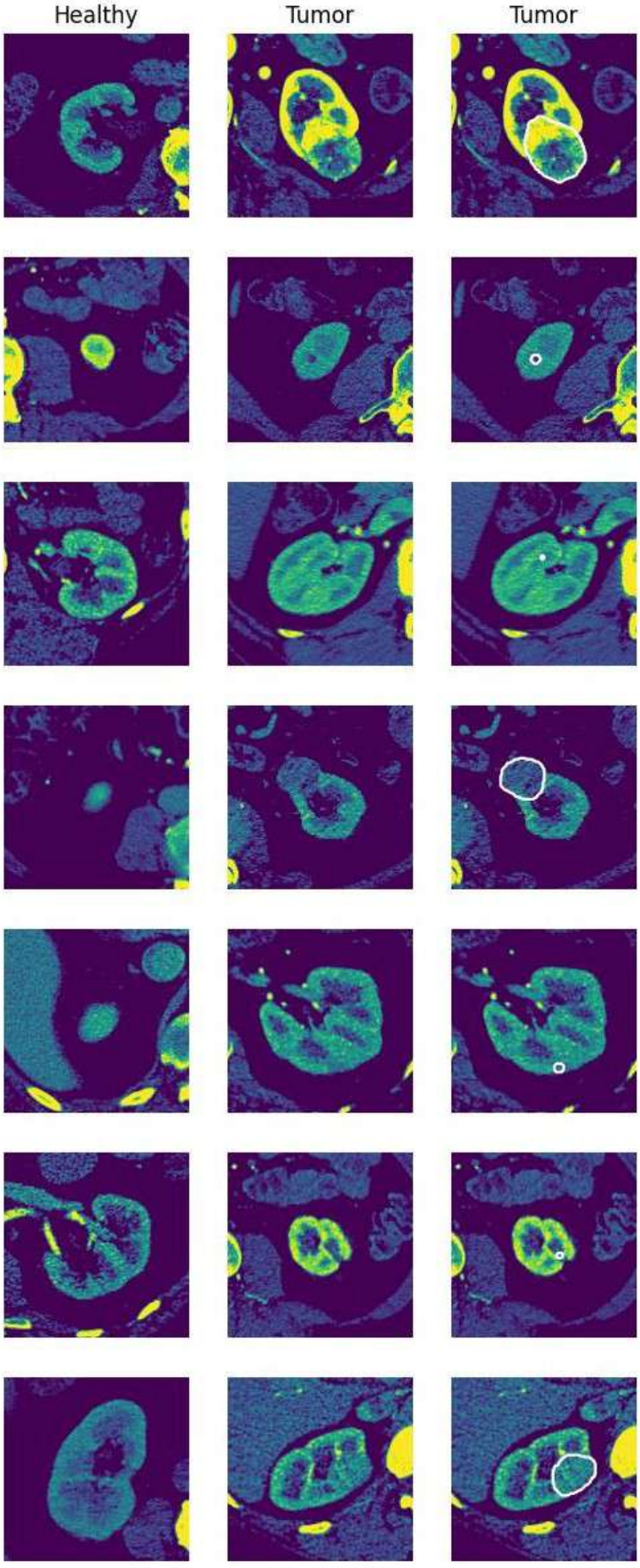

FIGURE 3.3: An illustrative example extracted from the prepared dataset, showcasing both tumor and healthy kidney samples.

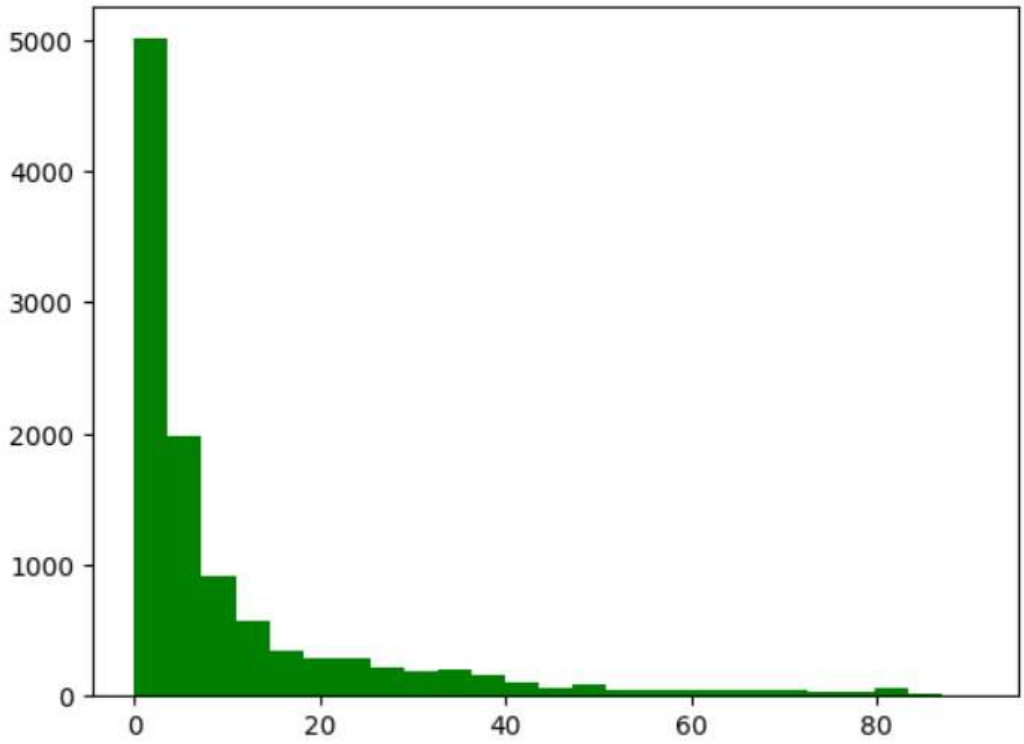

FIGURE 3.4: The tumor coverage within the image is measured by representing the percentage of the image on the x-axis and the number of examples on the y-axis.

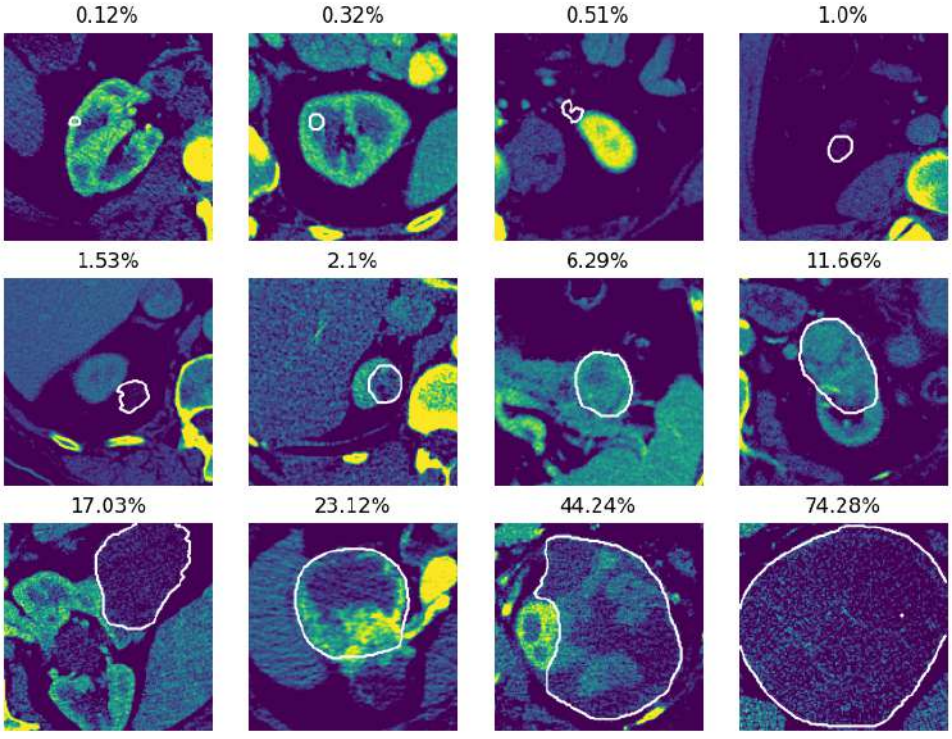

FIGURE 3.5: The area of tumor coverage within the image, represented as the percentage of the image.

Chapter 4

Experiments and results

Our study employed Kits21 dataset for training. The dataset was divided into different groups of data, and we will outline the details of this division. Furthermore, we will present a comprehensive comparison of the results obtained from various supervised classification models, which will highlight the strengths and weaknesses of each model. Finally, we will explore the most captivating aspect of our research: the results derived from a weakly supervised approach, incorporating visual representation.

4.1 Train data

The dataset was partitioned into distinct categories, with each category serving a specific purpose within our study. The splitting process ensured that each patient was assigned to only one split, eliminating any overlap between the training and test sets. This division of the dataset was implemented as follows:

- Training data : This subset constituted the majority of the dataset and was utilized to train our models. It played a crucial role in capturing patterns and relationships within the data. The training data accounted for 60% of the entire dataset.
- Validation data: To fine-tune the performance of our models during the training process, we allocated a portion of the dataset as validation data. This subset, comprising 20% of the dataset, allowed us to monitor the models' progress and make necessary adjustments.
- Test Data: The final portion of the dataset, accounting for 20%, was designated as the test data. This set remained completely unseen by our models during the training phase. We employed the test data to evaluate the models' overall performance, ensuring that they could generalize well to new, unseen instances.

Type of data	Number of images	Percent of images with tumor
Train	21816	28.96%
Validation	7510	28.50%
Test	7953	30.32%

TABLE 4.1: Detailed data split

4.2 Classification model results

A classification model is a type of supervised approach that requires both the image data and corresponding labels during training.

We conducted experiments using two different architectures, namely ResNet50 and ResNet101. Due to the difference in image dimensions, we were unable to leverage pretrained networks, as most pretrained models are designed for 3-dimensional images, while our dataset consists of 1-channel images. Additionally, we explored the impact of training data on our results. We pursued two approaches: firstly, employing every single image from the volume of the CT scan, and secondly, selecting every 5th image from the scan.

We selected the ResNet architecture due to its ease of training and its suitability as a foundational framework for various Class Activation Mapping (CAM) techniques. The concept of using a subset of images from the scan, rather than all of them, proved intriguing. Analogous to videos, where neighboring frames often exhibit high similarity, the same holds true for volumetric scans. Consequently, our objective was to direct the model's attention away from the general background and towards the organ and tumor regions of interest. By adopting this selective approach, we aimed to refine the model's focus and enhance its ability to discriminate between anatomical structures and pathological areas.

Default parameter values for both models are:

- learning rate is 1e-2,
- optimizer is Adam,
- batch size is 60,
- image train size is 150x150,
- decay rate is 0.1,
- decay epoch is 40.

These parameters were selected based on the results of the best performing experiments.

Based on the learning curve depicted in Figure 4.1, it is evident that the model begins to exhibit signs of overfitting after the 20th epoch. In general, the "ResNet101, all data" experiment performed better than the "ResNet50, part of data" experiment.

Experiment	Precision	Recall	F1-score
ResNet50, all data	75.69%	48.82%	59.35%
ResNet50, part of data	75.82%	57.36%	**65.31**%
ResNet101, all data	75.75%	**57.53%**	**65.39%**
ResNet101, part of data	**78.26%**	45.25%	57.35%
ResNet101, all data with augmentation	74.44%	55.45%	63.56%

TABLE 4.2: Experimental results of classification models

The experimental results indicate that ResNet50, being a shallower model, exhibits superior performance when confronted with a more diverse range of images, even when the available dataset is relatively smaller. On the other hand, ResNet101

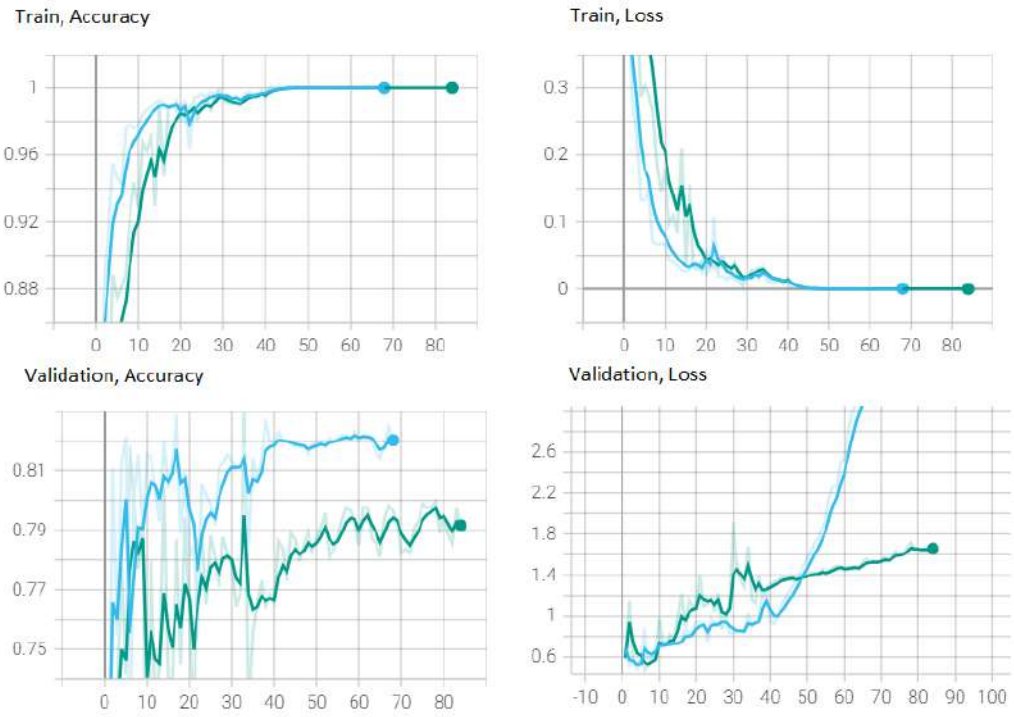

FIGURE 4.1: The learning curve plot illustrates the performance of two experiments: "ResNet101, all data" (represented by the blue line) and "ResNet50, part of data" (represented by the green line).

demonstrates a greater reliance on a larger quantity of images, even if they are relatively similar in nature. Notably, the outcomes obtained from ResNet50, trained on a reduced dataset, closely resemble those achieved by ResNet101 trained on the complete dataset. It is important to clarify that all the experiments were conducted using the same set of images over the entirety of the available data.

Due to the specific nature of the data and the presence of other organs, rotation and flipping augmentations cannot be applied. The noise in the data, even with minimal parameters, is too disruptive. Consequently, the only augmentation technique used was Gaussian blur, but it did not yield any improvements in the results, visualized in Table 4.2.

The confusion matrix for ResNet101, trained on the entire dataset, reveals that the model has a significant number of false negatives, indicating that it missed a considerable number of tumor images. This limitation can be attributed to the relatively small size of the available data, which may restrict the model's ability to effectively learn the underlying patterns. The findings emphasize the importance of augmenting the dataset with additional samples to improve the model's performance in detecting tumors.

Figure 4.2 illustrates the areas where the model's predictions deviate from the ground truth, highlighting the instances where incorrect predictions occur.

For all subsequent experiments, unless otherwise specified, we will use ResNet101 trained on the entire dataset. The computations and analyses in these experiments will be performed on the subset of correctly classified tumor images.

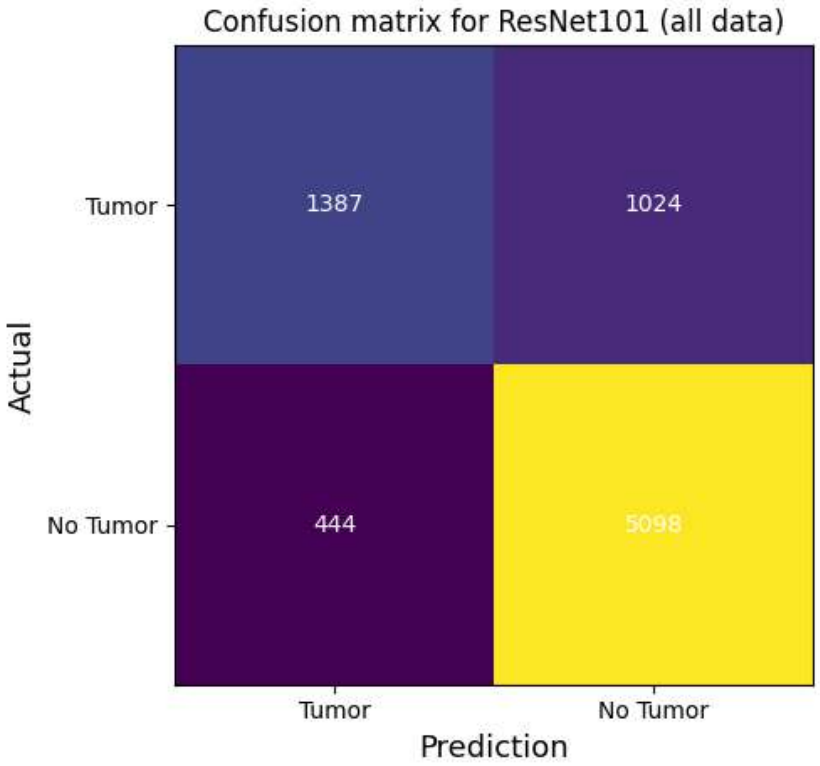

FIGURE 4.2: The Confusion Matrix of the outcomes of the ResNet101 model trained on the entire dataset.

4.3 Weakly-supervised models results

Initially, we conducted a preliminary evaluation of our pipeline to ensure its functionality and assess its performance on a simple and straightforward task. We introduced a black square in the top left corner of each image with a tumor class. This task served as an uncomplicated classification assignment, allowing us to verify the effectiveness of our unsupervised approach. The model demonstrated the ability to correctly classify images containing the square. In certain instances, our training images already contained a black background in that region, which compelled the model to focus on the tumor itself since it could not rely on the square's edges. This observation provided valuable insight into the model's attention towards the tumor region.

Subsequently, we proceeded with experiments using different approaches, namely Score-CAM, Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++. We carefully configured these experiments to determine the optimal threshold for each approach based on the validation set. The results of these experiments are presented in Table 4.3, showcasing the performance of each approach with their respective chosen thresholds.

Each of the approaches we employed in our experiments generates a heatmap as its output, with values ranging from 0 to 1.3. It is important to note that a threshold of 0 does not necessarily imply that the entire image is highlighted, but rather a substantial portion of it. Based on the results obtained, we can conclude that Grad-CAM and Score-CAM exhibit the most promising performance for our specific task. However, there is still room for improvement in these approaches, as well as in the

Approach	Threshold	IOU Validation	IOU Test
Grad-CAM	0.15	**0.367**	0.316
Grad-CAM++	0	0.303	0.251
Smooth Grad-CAM++	0	0.303	0.251
Score-CAM	0.15	**0.370**	**0.323**

TABLE 4.3: Experimental results for various CAM approaches

other methods explored. Visual representations of the results are presented in Figure N, providing a comprehensive overview of the heatmap outputs.

The presented Figure 4.3 showcases the results obtained from various unsupervised CAM methods. The ground truth region is depicted by the white line, while the blue line represents the prediction. The heatmap generated by each method provides insights into the focus and coverage of the predictions. In cases where a large portion of the image appears red, it indicates that the approach has bounded a significant area, but with some small values. The red color in the heatmap represents the prediction, and the presence of holes indicates that certain regions have been excluded from the prediction.

Examining the examples, we observe that in the first image, all the models correctly detected the tumor. However, in the second image, Grad-CAM exhibited the poorest performance, although it still captured approximately half of the tumor. The third example highlights that some methods focused either on the entire image or solely on the entire kidney, failing to provide precise tumor localization. Moving to the fourth example, we observe that Score-CAM demonstrated a strong focus on the tumor, while Grad-CAM failed to detect it entirely.

In the subsequent images, we find a large tumor in the fifth example, which was nearly fully detected by all the approaches. The sixth image exhibits a detected tumor, accompanied by a substantial area of prediction surrounding it. The seventh image showcases a small tumor that was successfully detected, although with a wider focus. Lastly, the eighth image features a sizable tumor that was accurately identified by all the employed methods.

Overall, the presented visual results demonstrate the performance and characteristics of different unsupervised CAM methods, highlighting their strengths and areas for improvement.

In Figure 4.4, it is observed that the model frequently detects the presence of the tumor, indicated by a non-empty intersection. However, the model's predictions often lack precision in terms of the exact tumor region. In some instances, the predicted area is too small compared to the actual size of the tumor. Conversely, there are cases where the model's focus on the tumor results in a larger predicted region.

The results of the experiment, visualized on Figure 4.5, demonstrate a similar trend to the Grad-CAM results. One of the large tumors exhibits a solid overlap with the gt (first image in the second row), while on another (last image in the third row) the heatmap generated by Score-CAM mostly misses the tumor. Locating small tumors proves to be challenging as the model focuses on different regions of the image.

In another experiment, we specifically tested on larger tumors. The rationale behind this approach is that detecting small tumors in a single image can be difficult. We anticipated improved results; however, it is worth noting that all the generated heatmaps often cover a larger area than the tumor itself. Therefore, when excluding

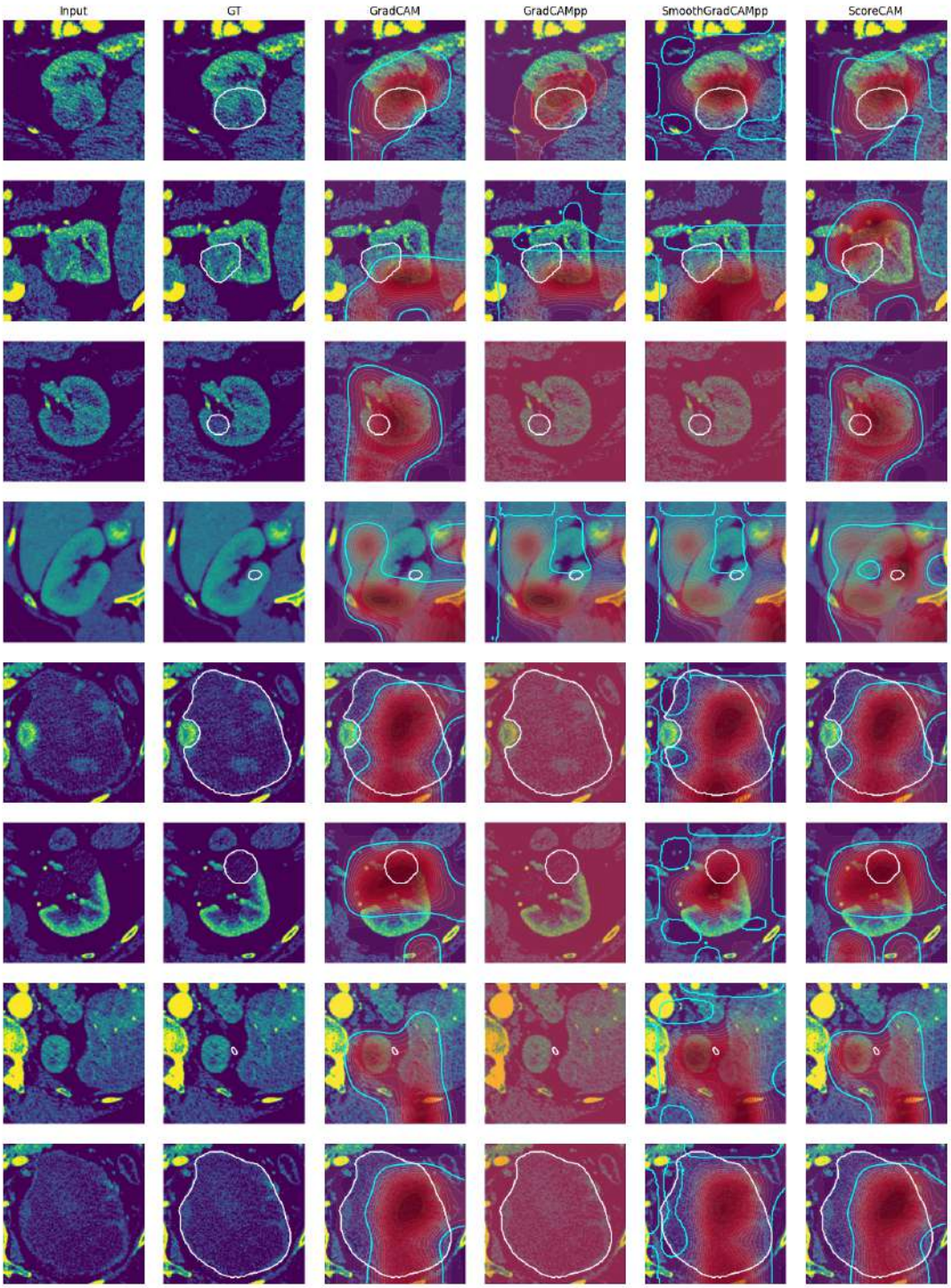

FIGURE 4.3: Comparison of Unsupervised CAM Methods: Ground Truth Region (White Line) vs. Prediction (Blue Line)

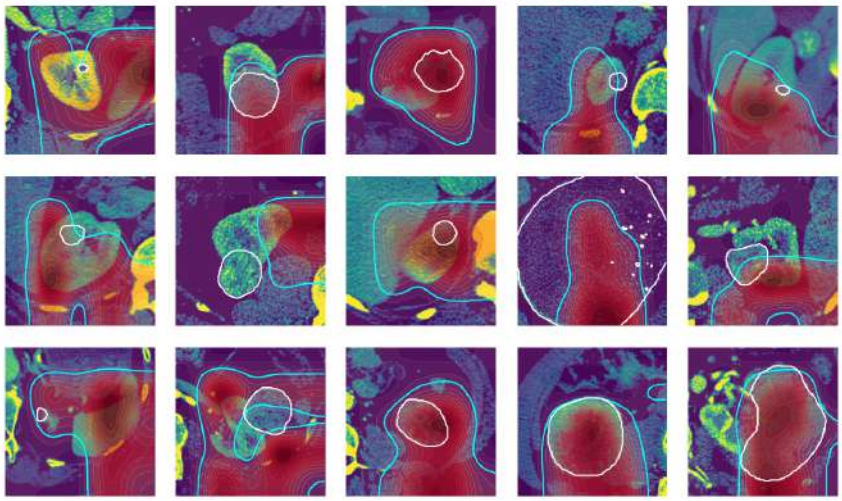

FIGURE 4.4: Visualization of Grad-CAM Results

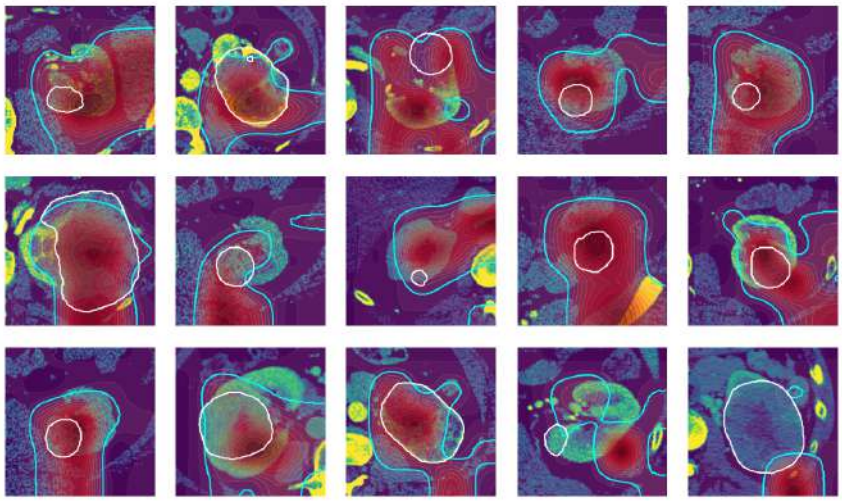

FIGURE 4.5: Visualization of Score-CAM Results

small tumors, whose regions are small and may result in a lower intersection over union (IOU) score, our predictions still encompass the tumor region. We selected tumors larger than 2% of the image for this experiment, and the results are presented in Table 3.

Approach	Threshold	IOU Validation	IOU Test
Grad-CAM	0.15	0.377	0.332
Score-CAM	0.1	0.382	0.338

TABLE 4.4: Experimental results for tumors with area > 2% of the image

The observed improvements in the model's performance were minimal, indicating that the issue with the heatmap persists, contrary to our initial assumption.

4.4 Comparison with Supervised Approach

We conducted a comparison between our weakly-supervised approach and a supervised approach using the U-Net model. Both models were trained on the same dataset consisting of images with and without tumors. Due to the fact that input data is single-channel images, we are unable to utilize pretrained models that are typically designed for multi-channel inputs. We used the original U-Net architecture, where the lowest feature map size was adjusted to 64x64x512, in contrast to the standard 30x30x1024 configuration. The evaluation scores are presented in Table 4.5, and the visual results are displayed in Figure 4.6.

Model	Test IoU
GradCAM	0.316
ScoreCAM	0.323
UNet	0.542

TABLE 4.5: Comparison of the results between the supervised and weakly-supervised models

From Figure 4.6, we observe that the U-Net model performs well overall, although it still struggles to detect small tumors. This difficulty in detecting small tumors is also evident in the results of the CAM approaches. The CAM methods generate larger heatmaps for small-sized tumors, resulting in some instances where the tumor is not detected, such as the second row in the figure. In contrast, the U-Net model produces more precise segmentation masks, as demonstrated in the last example.

One possible explanation for the U-Net's performance is that the classification model focuses solely on detecting tumors in the image, allowing it to find the tumor in at least one region and stop searching in other regions. In future work, we could explore incorporating a loss function that encourages the model to examine the entire heatmap, or explore alternative mechanisms to address this limitation.

The observed limitations in performance could potentially be addressed by modifying the model's configurations, such as increasing the number of convolutions, to enhance its ability to learn complex patterns. However, it is important to note that

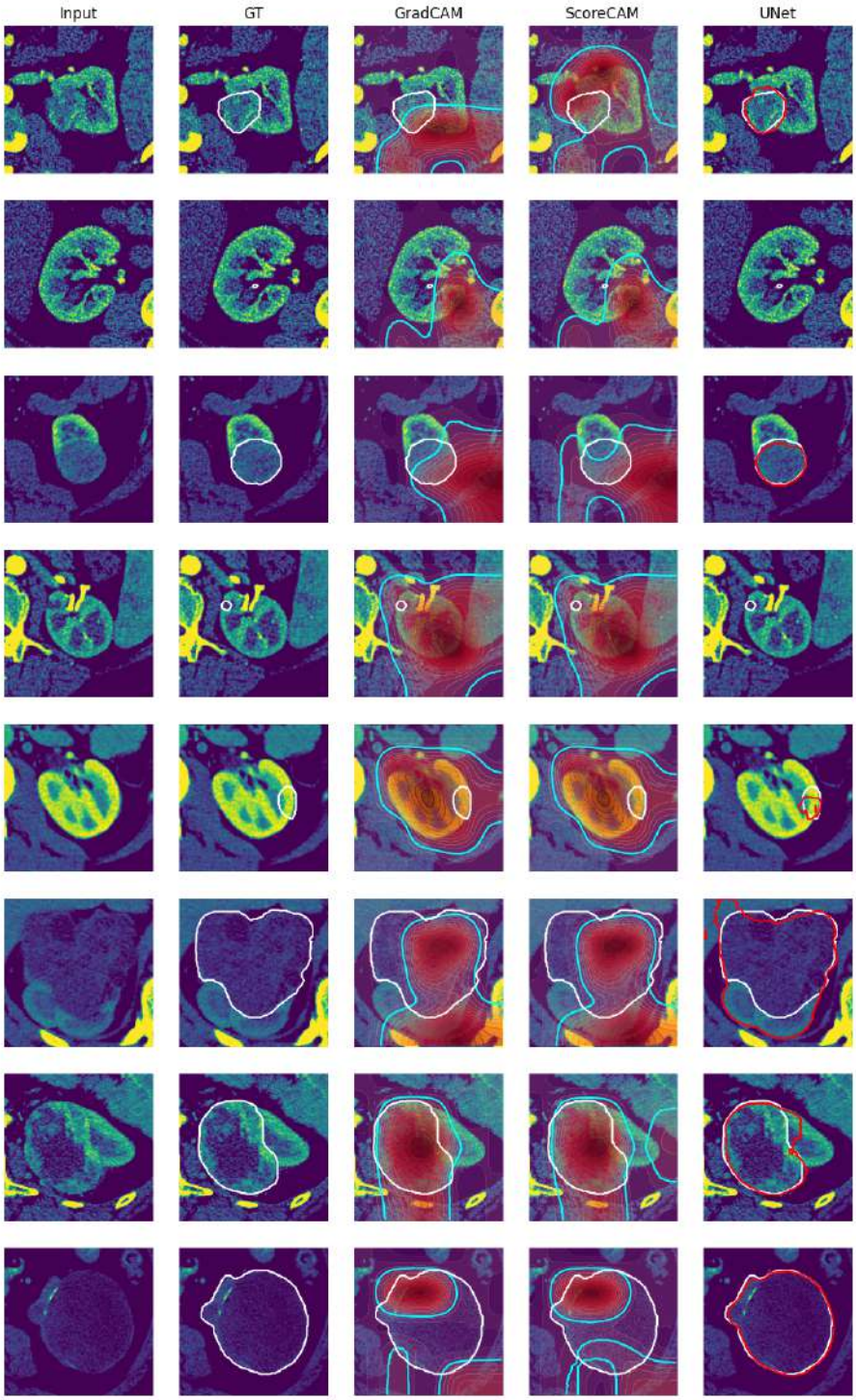

FIGURE 4.6: A comparison of the results between the supervised and weakly-supervised models

the lack of available data remains a significant challenge in achieving optimal results. The limited amount of data may slow down the model's ability to generalize effectively and capture the full range of tumor characteristics.

4.5 Hypothesis Results

In this chapter, we present the results obtained from testing the hypotheses formulated at the beginning of our research. The hypotheses were designed to explore various aspects of weakly-supervised tumor segmentation and compare it with supervised approaches. The following sections provide a summary of the findings and their implications.

Hypothesis 1: Weakly-supervised approaches have the potential for tumor segmentation and can achieve promising results. Result: Our experiments confirm that weakly-supervised approaches have the potential to effectively segment tumors. The obtained results demonstrate the feasibility of utilizing weakly-supervised learning methods for tumor segmentation tasks.

Hypothesis 2: Score-CAM, a specific CAM approach, can outperform Grad-CAM approaches in terms of performance. Result: Our analysis reveals that Score-CAM indeed outperforms other Grad-CAM approaches in terms of performance. It consistently demonstrates superior segmentation accuracy, highlighting its effectiveness as a CAM technique for tumor segmentation. However, it should be noted that Grad-CAM approach is also competitive and show comparable performance in certain configurations.

Hypothesis 3: Given the limited amount of available data, employing shallower network architectures holds potential for yielding improved results in our task. Result: Our findings suggest that these architectures, which prioritize the learning of high-level features, exhibit promising performance in capturing important tumor characteristics. However, it is worth noting that in our specific case, deeper models demonstrate superior performance.

Hypothesis 4: Supervised approaches will outperform weakly-supervised approaches in terms of segmentation accuracy, as they directly learn the correct segmentation. Result: The experiments confirm that supervised approaches generally outperform weakly-supervised approaches in terms of segmentation accuracy. However, the difference in performance between the two approaches is relatively small, with an IoU score difference of 0.23. This suggests that further improvements can be made to enhance the performance of weakly-supervised approaches in tumor segmentation.

The results of our hypothesis testing provide valuable insights into the strengths and limitations of weakly-supervised tumor segmentation approaches. While weakly-supervised methods show promise and can achieve competitive results, there is still room for improvement. Future work should focus on refining weakly-supervised techniques to bridge the performance gap with supervised approaches.

Chapter 5

Conclusion

This research presents a comprehensive investigation into weakly-supervised tumor segmentation using a pipeline that incorporates a ResNet101 classification model and the Score-CAM approach. Through the use of higher activations to identify tumor regions, the proposed pipeline showcases the benefits of learning from a full dataset with image-level labels. The inclusion of the Score-CAM technique yields a noteworthy IoU score of 0.323, underscoring the effectiveness of the weakly supervised approach in tumor segmentation tasks. The research provides detailed insights into implementation strategies and thorough qualitative and quantitative evaluations to support the findings.

This approach has the potential to assist labellers by highlighting areas of high interest, thereby labellers can focus their attention on these specific regions, making the labelling task more efficient and effective.

In terms of future research directions, there are several important steps to consider. Firstly, it is crucial to focus on enhancing the performance of Grad-CAM++ and Smooth Grad-CAM++, as well as exploring the potential of other CAM approaches.

Additionally, an alternative direction to pursue involves incorporating 3D Score-CAM and leveraging voxel-based training. By using the sequential information present in the input data, there is a potential for improved model performance.

Furthermore, an interesting avenue to explore is the combination of LSTM and attention mechanisms to feed the model with voxel slices in a sequential manner. This approach has the potential to capture temporal dependencies and further enhance the model's segmentation capabilities.

These future steps hold promise for advancing the current research and potentially improving the accuracy and effectiveness of tumor segmentation.

For those interested in the code used in this research, it will be made available via email at zakharchenko@ucu.edu.ua. Moreover, the dataset used in this study is open and accessible for further exploration.

Bibliography

Cai, Jinzheng et al. (2017). *Improving Deep Pancreas Segmentation in CT and MRI Images via Recurrent Neural Contextual Learning and Direct Loss Function*. arXiv: 1707.04912 [cs.CV].

Chattopadhay, Aditya et al. (2018). "Grad-CAM: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. DOI: 10.1109/wacv.2018.00097. URL: https://doi.org/10.1109\%2Fwacv.2018.00097.

CT scan (2023). URL: https://en.wikipedia.org/wiki/CT_scan.

Feng, Xinyang et al. (2017). "Discriminative Localization in CNNs for Weakly-Supervised Segmentation of Pulmonary Nodules". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Springer International Publishing, pp. 568–576. DOI: 10.1007/978-3-319-66179-7_65. URL: https://doi.org/10.1007\%2F978-3-319-66179-7_65.

He, Kaiming et al. (2015). *Deep Residual Learning for Image Recognition*. arXiv: 1512.03385 [cs.CV].

Heller, Nicholas et al. (2020). *The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes*. arXiv: 1904.00445 [q-bio.QM].

Hollo, Kaspar (2019). *Exploring the Value of Weakly-Supervised Deep Learning Approaches for Artefact Segmentation in Brightfield Microscopy Images*. URL: https://comserv.cs.ut.ee/home/files/hollo_softwareengineering_2021.pdf?study=ATILoputoo&reference=FB62B954025B6EA4F56FA7C1AF0A89C914A957F0.

Kidney Cancer (2023). URL: https://www.cancer.net/cancer-types/kidney-cancer/introduction.

Liu, Yang et al. (2022). *Mixed-UNet: Refined Class Activation Mapping for Weakly-Supervised Semantic Segmentation with Multi-scale Inference*. arXiv: 2205.04227 [eess.IV].

Omeiza, Daniel et al. (2019). *Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models*. arXiv: 1908.01224 [cs.CV].

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. arXiv: 1505.04597 [cs.CV].

Ryan, Sarah et al. (2020). *Cluster Activation Mapping with Applications to Medical Imaging*. arXiv: 2010.04794 [cs.CV].

Selvaraju, Ramprasaath R. et al. (2019). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2, pp. 336–359. DOI: 10.1007/s11263-019-01228-7. URL: https://doi.org/10.1007\%2Fs11263-019-01228-7.

Simple Kidney Cysts (2019). URL: https://www.niddk.nih.gov/health-information/kidney-disease/simple-kidney-cysts.

Wang, Haofan et al. (2020). *Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks*. arXiv: 1910.01279 [cs.CV].

Yang, Guanyu et al. (2020). "Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal CTA images". In: *BMC Medical Imaging*

20.1, p. 37. ISSN: 1471-2342. DOI: 10.1186/s12880-020-00435-w. URL: https://doi.org/10.1186/s12880-020-00435-w.

Zhou, Bolei et al. (2015). *Learning Deep Features for Discriminative Localization*. arXiv: 1512.04150 [cs.CV].