Disaster Recognition Through Image Captioning Features and Shifted Attention

Submitted to
Graduate School of Information Science and Technology
Osaka University

September 2024

Narongthat THANYAWET

Thesis Committee:

Prof. Yuki Uranishi (Osaka University)
Prof. Tatsuhiro Tsuchiya (Osaka University)
Prof. Yoshinobu Kawahara (Osaka University)
Assoc. Prof. Shizuka Shirai (Osaka University)

List of Publications

Journals

1. N. Thanyawet, P. Ratsamee, Y. Uranishi, M. Kobayashi and H. Takemura. Identifying Disaster Regions in Images Through Attention Shifting with a Retarget Network. *IEEE Acess*, 2024.

International Conferences

Peer-reviewed Papers

1. N. Thanyawet, P. Ratsamee, Y. Uranishi, and H. Takemura. Abnormal Scene Classification using Image Captioning Technique: A Landslide Case Study, *IEEE Conference on Pattern Recognition Systems (ICPRS)*, 1–7, Jul. 2023.

2. T. Boonchob, N. Tuaycharoen, S. Limpeeticharoenchot, and N. Thanyawet. Job-Candidate Classifying and Ranking System-Based Machine Learning Method, *IEEE International Computer Science and Engineering Conference (ICSEC)*, 94–99, Dec. 2022.

National Conferences

1. N. Thanyawet, P. Ratsamee, Y. Uranishi, and K. Arai. Using Detective Network for Anomaly Detection in Images, *The 22th International Symposium on Automation and Robotics in Construction*, October 8-10, 2024 Ibaraki, Japan.

Abstract

In the wake of escalating natural disasters, timely and precise recognition of affected areas has become imperative for efficient disaster management and mitigation. Traditional image processing techniques often fall short in identifying subtle nuances within disaster-stricken regions due to their propensity to highlight prominent features, thereby overlooking critical details. This dissertation presents a method that utilized image captioning features along with adaptive attention mechanisms to improve the recognition of disaster-affected regions.

The author proposes a novel methodology that integrates image captioning with a custom-developed attention-shifting algorithm designed to dynamically refocus the model on less conspicuous yet essential elements within images. By leveraging the inherent strengths of Vision Encoder-Decoder (VED) models, along with innovative optimal masking strategies, we enable the system to discern and articulate the specifics of disaster impacts in diverse imaging conditions, from satellite to ground-level perspectives.

The research methodology includes the rigorous training and evaluation of the model using extensive datasets comprising side-view, aerial, and shipborne images of disaster scenes. The model's performance is assessed against standard metrics, demonstrating a significant leap in accuracy and contextual relevance of the generated captions.

The empirical results underscore the superiority of our approach over conventional image captioning models, exhibiting enhanced detection capabilities with accuracies exceeding 91% for landslide detection from side-view image captions and 87.5% for shipborne view detection. These figures not only reflect the technical prowess of the system but also its practical applicability in real-world disaster assessment scenarios.

This work carries profound implications for the field of disaster management. By augmenting the quality and reliability of disaster region identification, our framework facilitates more informed decision-making in allocating resources for relief efforts. Additionally, the adaptive nature of the model paves the way for its application across a spectrum of environmental monitoring and emergency response tasks, heralding a new era of AI-enabled disaster management tools. Future research avenues include scaling the model to encompass a broader range of disaster types and integrating real-time data for swift, actionable insights during crisis events.

Acknowledgments

I have joined the Integrated Media Environment Lab (Takemura Lab) for my doctoral course at Osaka University. At that time, it was quite challenging due to the COVID-19 situation, which prevented me from coming to Japan in the first semester of my doctoral course. Since I enrolled in this new journey, I have received a lot of support in various aspects from our laboratory members. I would like to thank the following people who supported me during this wonderful journey.

First, I would like to thank my supervisor, Prof. Haruo Takemura, for all his support in my scholarship and Ph.D. life. Moreover, I would like to thank Prof. Yuki Uranishi for supporting and continuing the IME Lab in my last semester and for providing good comments all the time.

I received a lot of comments that significantly improved my work, and I would like to thank Assoc. Prof. Photchara Ratsamee for providing the opportunity for this journey.

Furthermore, I would like to thank my dissertation committee for providing their valuable time to me with good comments and feedback.

Special thanks to Sysmex Corporation for the research grant and special internship which provided me with valuable experience in my journey. Additionally, thanks to the Thai Student Community in Japan (TSAJ) for their encouragement and support during my toughest periods.

Finally, I would like to thank my family, who always encouraged me. I would like to extend a huge thank you to Dr. Ronnakorn Vaiyawuth for always supporting and coaching me during the toughest times of my life. Last, Pruetiwan Asawakowitkorn and Suchaya Naruetanapakorn for always check and encourage me in the worst day.

Narongthat Thanyawet
Osaka University
December 2024

Contents

List of Tables

List of Figures

List of Acronyms

CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
VED	Vision Encoder Decoder
ViT	Vision Transformers
BERT	Bidirectional Encoder Representations from Transformers
Conv	Convolutional
FC	Fully Connected
UAV	Unmanned Aerial Vehicle
NLP	Natural Language Processing
GPU	Graphics processing unit

Chapter 1

Introduction

Computer vision is increasingly being applied to address various global challenges (Lis et al., 2019; Kamoi et al., 2021; Ohgushi et al., 2020; Di Biase et al., 2021; Chan et al., 2021). A particularly critical application is the detection of disaster-affected areas in images, which may be sourced from Unmanned Aerial Vehicles (UAVs), helicopters, or ships.

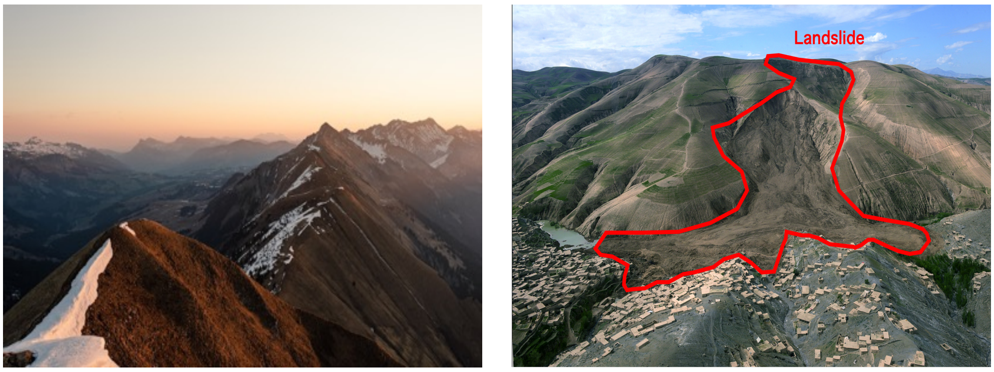
Disasters often involve the displacement of natural elements, such as in landslides, where soil or rocks shift from higher elevations, and floods, which occur when water overflows onto impermeable surfaces or within urban environments.

One major challenge is that the images used for detection typically cover vast areas, making the disaster-affected regions relatively small or not immediately apparent. Moreover, the visual similarity of landslides to regular soil and flooded areas to typical bodies of water complicates the use of machine learning methods for accurate detection.

1.1 Disaster Detection

Nowadays, object detection in image has become a common challenge in various fields (Meena et al., 2022; Li et al., 2022; Can et al., 2019). Computer vision techniques are used in disaster investigations to evaluate the damaged areas before rescuing victims or reconstructing the affected regions. For these purposes, aerial investigations have become the most popular method to gather information from an aerial view, using UAVs or helicopters to collect images from a bird's-eye view to obtain an overview of the disaster situation. However, it is difficult for machine learning models to detect disaster regions from images taken from an aerial perspective (Ofli et al., 2021, 2022).

A disaster scene is defined as natural objects misplaced from their common situations. For instance, in Figure 1.1, which represents a non-disaster (left) and disaster (right) scene, the non-disaster image shows a mountain covered by soil and rock, which is a regular situation since the soil and rocks typically cover the mountain. However, in the right image, the soil within



Non-disaster scene

Disaster scene (landslide)

Figure 1.1: Difficulty in classification of non-disaster and disaster scene due to visual similarity.

the red boundary appears misplaced and has slid down from the hill, characterizing this as a disaster scene. It involves the same materials-soil, rocks, grass, etc.–but in an irregular place or situation. For this reason, objects in irregular positions can lead to what are termed *abnormal objects*; however, detecting such disaster or abnormal cases is quite challenging because the disaster regions still comprise natural objects, just in different positions.

In a similar vein, flooding presents the same challenges as landslide disasters, albeit with slight differences. Flooding occurs when water is misplaced into household areas or other regions outside of waterways (e.g., rivers, canals, ponds, reservoirs, etc.), which makes detecting floods easier than landslides in some aspects. However, detecting water bodies can be difficult (Hernández et al., 2022) from certain viewpoints, such as when reflections from the water surface mirror other objects.

On the other hand, wildfires are difficult to detect from certain perspectives, such as scenes obscured by many objects among the smoke. However, from an aerial view, wildfires are easier to identify due to the color differences in the images and the surrounding objects.

Therefore, detecting natural disasters in computer vision presents a significant challenge, particularly when the scenes are similar. Many techniques in computer vision and image processing, including pre-processing, neural networks, or deep learning, do not significantly enhance disaster detection (Ofli et al., 2021, 2022; Hernández et al., 2022). Moreover, in computer vision, in Figure 1.2 machine learning techniques that use CNNs are based on localized pixel features and tend to focus on centering or highlighting prominent objects rather than other parts of the scene, which might contain significant but tiny

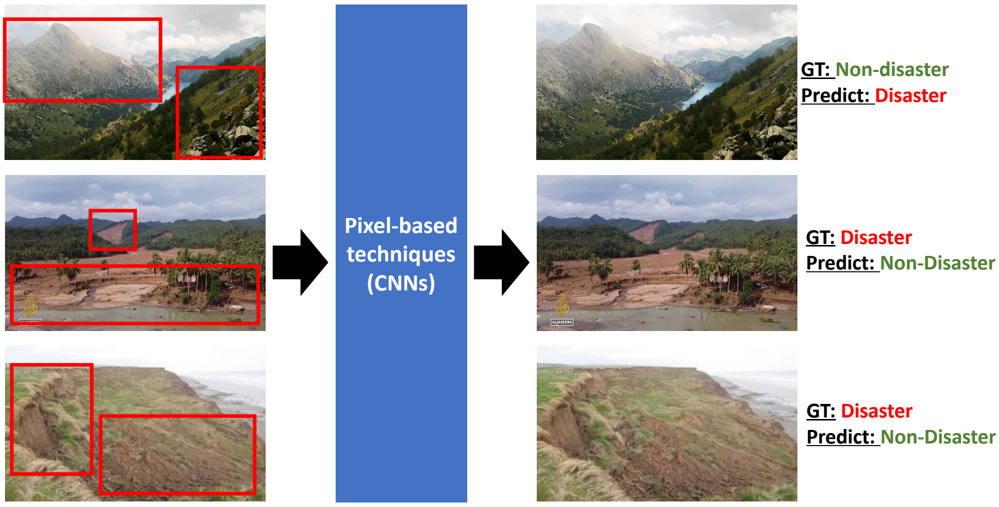


Figure 1.2: The challenge for pixel-based techniques (CNNs) (Meena et al., 2022; Soares et al., 2020; Liu et al., 2020) to classify disaster scene.

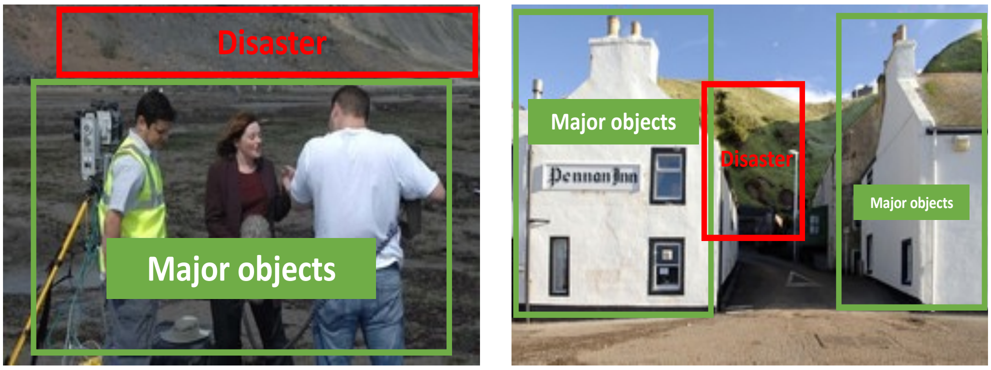
regions.

Moreover, pixel-based techniques analyze the surrounding pixels to classify the scene into various categories. These methods typically use the color, intensity, and texture information of neighboring pixels to determine the class of each pixel in the image. However, conventional pixel-based methods have certain limitations, particularly when it comes to complex or atypical scenarios. In my case study, the objects that must be detected are natural objects displaced from their usual locations due to a disaster.

Conventional techniques may struggle to accurately classify these objects because they are designed to recognize and categorize typical scenes. This leads to a critical challenge: these methods might incorrectly classify the displaced natural objects. The potential consequences of such misclassification are significant, as they could either mistake them for ordinary, non-disaster-related natural objects or fail to recognize them as indicators of a disaster. This misclassification occurs because traditional approaches often do not account for the contextual and situational anomalies present in disaster scenarios.

1.2 Attention in Computer Vision

Computer vision techniques (Meena et al., 2022; Soares et al., 2020; Liu et al., 2020) tend to use the features from neighboring pixels' color, intensity,



The main object is human in foreground

The disaster object is small compared with surrounding

Figure 1.3: Difficulty in detection the disaster region with the huge major objects in the scene.

and texture information to detect or classify the scenes. However, detecting disaster-related objects poses a significant challenge. This difficulty arises because natural objects not associated with disasters often have visual features similar to those that are not.

Figure 1.3 shows that the target object may not be the model's primary focus since it might be in the background, where other objects are more prominent and attract more attention than the target region. Moreover, the correct figure in 1.3 demonstrates that the disaster object is not the main focus and appears small compared to the entire scene.

Conventional methods in the field of computer vision typically utilize surrounding pixels to extract features in order to achieve their objectives. However, these computer vision models often prioritize attention or focus on the scene's center. The models often concentrate primarily on foreground objects or the main focal points. Recently, the attention mechanism from the transformers architecture (Vaswani et al., 2017) has gained popularity for addressing challenges in natural language processing and computer vision. This technique employs the attention method to better focus on target objects. However, despite its advancements, the attention layer still tends to prioritize foreground objects or the main objects, similar to conventional techniques, as shown in Figure 1.3.

In this research, the author utilizes information from pixel-based techniques and image captioning, which provides more meaningful features than conventional methods. Moreover, the challenge of model focus-where attention is given to primary objects or target regions-poses a significant problem,

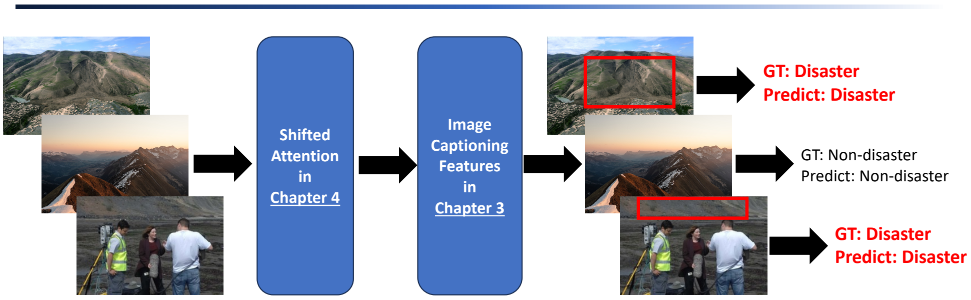


Figure 1.4: The research goal to detect the disaster from the image.

especially when the disaster region is too small to detect. The proposed architecture incorporates re-targeted attention to address this issue, as represented in Figure 1.4.

1.3 Statement of the Problem

- Detecting disaster scenes using simple pixel-based information, such as color, texture, or intensity, presents significant challenges. It is difficult to distinguish disaster-related objects from non-disaster objects in a natural scene using these features alone. The visual similarities between natural objects that are part of a disaster and those that are not can lead to misclassification, making it hard for conventional pixel-based methods to identify disaster scenarios accurately.

- Disaster-related objects might be situated in the background, causing the model to focus on other, more prominent objects instead of the primary target, the disaster event. Prioritizing the focus of a machine learning model on these disaster events is one of the most difficult challenges. Ensuring that the model accurately identifies and prioritizes the relevant features associated with disaster scenarios, despite their often subtle or background presence, requires advanced techniques and careful tuning.

1.4 Research Questions

- The first question is: How can we extract more meaningful features from images? Pixel-based techniques often fail to accurately classify scenes due to a lack of sufficient information and features, which can lead to misunderstandings between non-disaster objects and disaster objects.

To improve the accuracy of scene classification, it is essential to develop methods that can capture and utilize more meaningful and contextually relevant features from the images.

- The second question is: How can we detect disaster-related objects that differ from environmental objects, given that these disaster objects are natural objects that have been misplaced in the environment? This challenge arises because disaster-related objects often have similar visual characteristics to non-disaster objects, making it difficult to distinguish between them using conventional detection methods.
- The third question is: How can we prioritize the model's attention? Computer vision models that utilize the features extracted from CNN layers tend to focus on the significant objects in a scene. However, a scene can contain multiple objects, and ensuring that the model appropriately prioritizes its attention is crucial. Developing techniques to guide the model's focus towards the most relevant objects, especially in complex scenes with multiple elements, is a significant challenge.
- The fourth question is: How can we select the optimal mask to shift attention? Utilizing masked regions to direct the attention of a machine-learning model is a viable approach. However, the masked regions can vary significantly in terms of aspects such as shape and location. Determining the optimal characteristics of these masks to effectively shift the model's attention to the most relevant areas is a complex task that requires careful consideration and experimentation.

1.5 Philosophy

Considering the research questions from the previous section, the author has established guiding philosophies to enhance the robustness of disaster detection. These philosophies encompass four main topics as follows in Figure 1.5.

First, meaningful features from the images should not be limited to pixel-based information but should include explainable features in human languages, such as captions. Individuals may perceive and describe the same image differently in each scene based on their unique experiences. For this reason, captions generated from images become a valuable feature for detection, providing contextual and semantic information that pixel-based methods alone may miss.

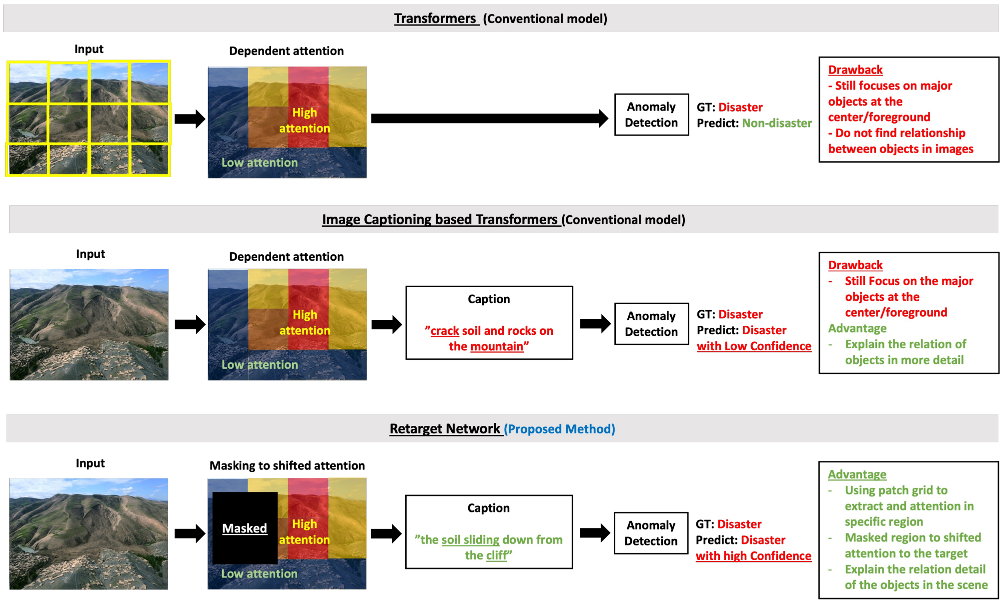

Transformers (Conventional model)
Input
Dependent attention
High attention
Low attention
Anomaly Detection
GT: Disaster
Predict: Non-disaster
Drawback
- Still focuses on major objects at the center/foreground
- Do not find relationship between objects in images
Image Captioning based Transformers (Conventional model)
Input
Dependent attention
High attention
Low attention
Caption
"crack soil and rocks on the mountain"
Anomaly Detection
GT: Disaster
Predict: Disaster with Low Confidence
Drawback
- Still Focus on the major objects at the center/foreground
Advantage
- Explain the relation of objects in more detail
Retarget Network (Proposed Method)
Input
Masking to shifted attention
Masked
High attention
Low attention
Caption
"the soil sliding down from the cliff"
Anomaly Detection
GT: Disaster
Predict: Disaster with high Confidence
Advantage
- Using patch grid to extract and attention in specific region
- Masked region to shifted attention to the target
- Explain the relation detail of the objects in the scene

Figure 1.5: Philosophies of research

Second, disaster objects in the environment often resemble non-disaster objects. Captions can provide detailed explanations, offering more meaningful features that capture the action-related aspects of these objects. Natural objects in normal situations typically do not exhibit these action features, allowing for effective classification based on this information. By leveraging the additional context provided by captions, it becomes possible to distinguish disaster objects from their non-disaster counterparts.

Third, conventional computer vision techniques focus on the center of the scene, major objects in the foreground, or large objects in the images. To effectively detect the target object, such as a disaster region, it is necessary to use masked images to re-target the detection towards the significant region. By applying masks during the initial attention phase, the model's focus can be directed away from less relevant areas and towards the critical regions that indicate the presence of a disaster.

Fourth, masking the image involves various parameters, such as the shape and location of the masked regions. The model would be trained using attention captioning techniques to optimize these masked regions. This approach allows the model to adjust the masks to the optimal size and position for the relevant region. By fine-tuning the masked areas during training, the model can better focus on the suitable regions, improving the overall accuracy and effectiveness of disaster detection.

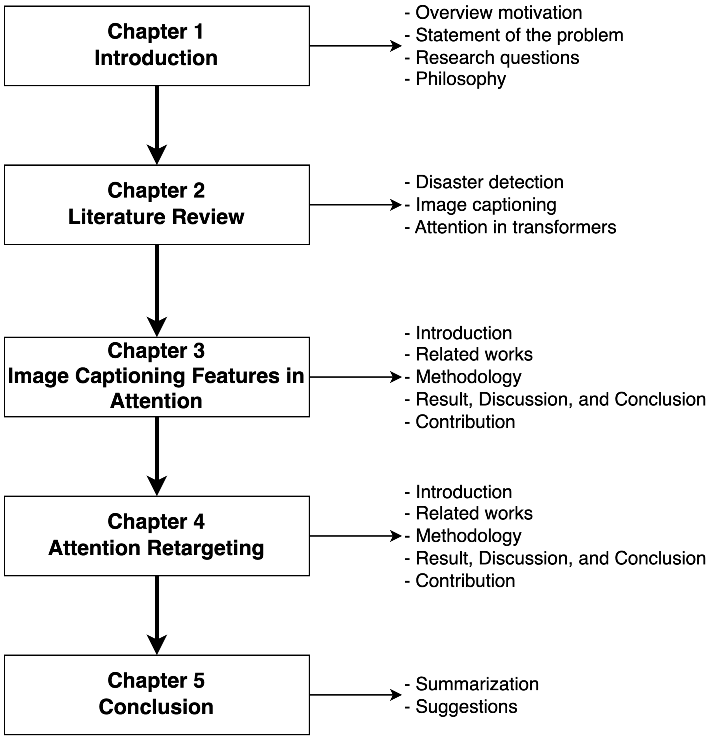


Figure 1.6: Dissertation overview

1.6 Outline of the Dissertation

The overall content of this dissertation is represented in Figure 1.6. The dissertation consists of five chapters, which are briefly described as follows:

1. Chapter 1: Introduction. This part begins with the background of disaster detection, traditionally in the computer vision field, and focuses on computer vision studies. Moreover, this chapter contains the problem statement, research questions, philosophy, and contribution of the proposed method.

2. Chapter 2: Literature Review. This part reviews previous studies on disaster detection, developments in computer vision such as Convolutional Neural Networks (CNNs), Image Captioning, and the *Attention* mechanism from Transformer architectures.

3. Chapter 3: Image Captioning Features in Attention. This part will explain the approach to obtaining meaningful information and features from the image; *Image Caption Features* could provide more detailed actions from the scene situation than conventional methods.

4. Chapter 4: Shifted Attention. This part will explain the novel approach to retarget the model in order to detect target regions that are in the background or quite small in the scene.

5. Chapter 5: Conclusion. This chapter summarizes the major findings, contributions, and suggestions for future work.

Chapter 2

Related Work

This chapter presents related works on three main aspects crucial for this dissertation's core development: disaster detection, image captioning, and attention mechanisms in transformers. Each area plays a significant role in advancing the methodologies and approaches discussed in this research.

2.1 Disaster Detection

Nowadays, computer vision is employed in various challenges to automate detection (Lis et al., 2019; Ofli et al., 2021; Krizhevsky et al., 2012; Johnson et al., 2016). In disaster events, monocular investigation from aerial imagery is used to detect disaster regions for planning and recovery efforts (Tantanee et al., 2018). Images captured from aerial investigations by Unmanned Aerial Vehicles (UAVs) or helicopters are processed to identify disaster regions using various computer vision methods, such as classification, detection, and segmentation.

The challenge in disaster detection arises because the characteristics of disaster-related objects are often similar to those of natural objects. For instance, a landslide involves soil and rocks sliding down from high ground and causing damage to constructions. It is difficult to distinguish between ordinary soil and rocks and a landslide disaster, especially if the event occurs in a countryside area. This similarity makes it challenging for conventional detection methods to identify and classify such events accurately.

(Ofli et al., 2021) and (Ofli et al., 2022) use conventional techniques to detect disaster regions from side-view images. The result using ResNet50, illustrated in Figure 2.1, is one of the most popular methods for classification challenges. However, there are limitations when detecting landslide disasters in scenes using the architecture of CNNs that rely solely on pixel features. In the case of landslides, disaster-related objects often resemble natural objects, making it quite challenging for these methods to detect and differentiate them accurately.

Since pixel-based techniques receive limited features from images' color, texture, and intensity, more is needed for classifying disaster regions. There-

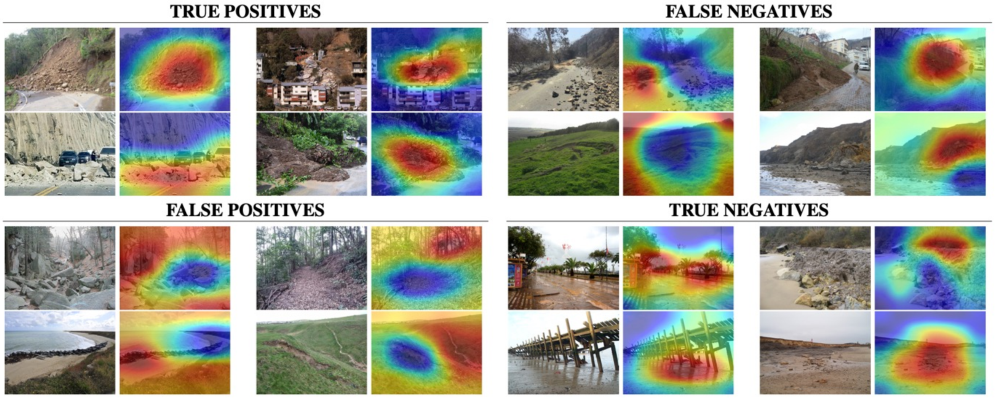


Figure 2.1: Confusion metric from landslide classification using ResNet50 (Ofli et al., 2021).

fore, using a model incorporating more features can improve the accuracy of disaster detection. In this case, captions from the images can provide more meaningful information, enriching the feature set used for disaster classification. The model can better differentiate between disaster-related objects and similar natural objects by leveraging the descriptive context provided by image captions.

2.2 Image Captioning

Recently, (Johnson et al., 2016) established an image captioning method to explain the situation within a scene. This method consists of two parts: the image encoder, which uses Convolutional Neural Networks (CNNs), and the caption generator, which employs Recurrent Neural Networks (RNNs).

Image captioning provides more detailed information about the content of images by generating sequential labels rather than the single labels used in classification methods. (Johnson et al., 2016) demonstrates state-of-the-art performance in extracting detailed information from images through dense image captioning, as represented in Figure 2.2. This approach yields meaningful features from the image by generating descriptive captions.

Image captioning utilizes an encoder part from CNNs to extract features and generates related captions from these encoded features using RNNs. This method provides more detailed features within the image than the single-word labels generated by image classification or detection methods. Therefore, combining image captioning with pixel-based conventional techniques presents a promising approach to addressing the disaster detection challenge. Integrating

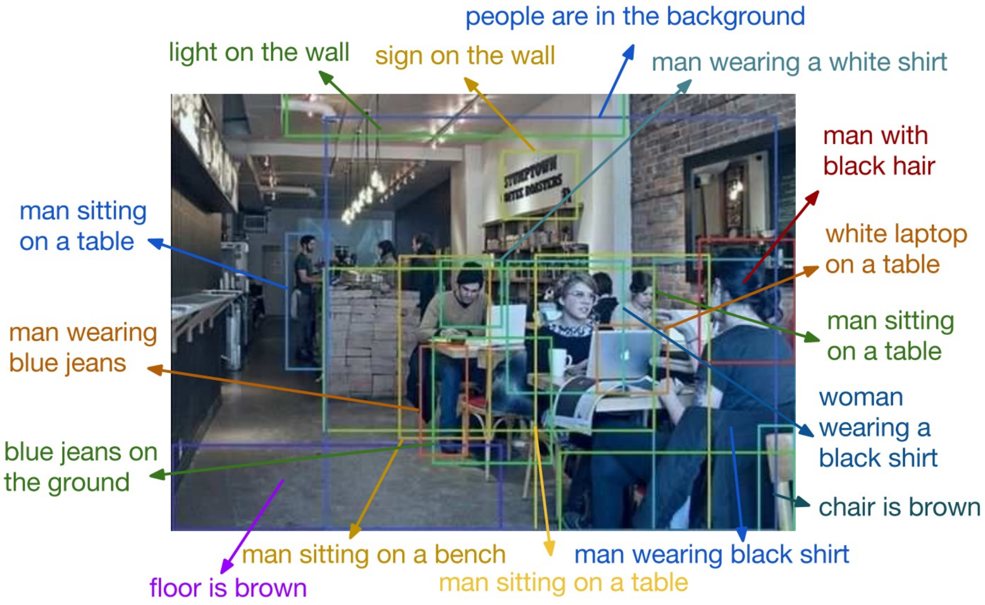


Figure 2.2: Dense image captioning (Johnson et al., 2016).

the rich, descriptive information from image captions with traditional pixel-based methods makes it possible to achieve more accurate and robust disaster detection.

2.3 Attention in Transformers

Computer vision techniques tend to detect objects located in the middle, foreground, or those that are large within a scene. This detection characteristic can lead to misdetection if the primary target objects are in the background or are very small within the scene. As a result, critical disaster-related objects might be overlooked, which can compromise the effectiveness of the detection process.

More recently, (Vaswani et al., 2017) established the state-of-the-art transformers architecture to address natural language processing challenges using the sequence-to-sequence method. Moreover, in the field of computer vision, transformer techniques have been applied as well. (Dosovitskiy et al., 2020) used the encoder block of transformers to extract valuable features through attention mechanisms. These features are then passed through a Multi-Layer Perceptron (MLP) to classify the image. This approach leverages the power of attention in transformers to enhance feature extraction and improve clas-

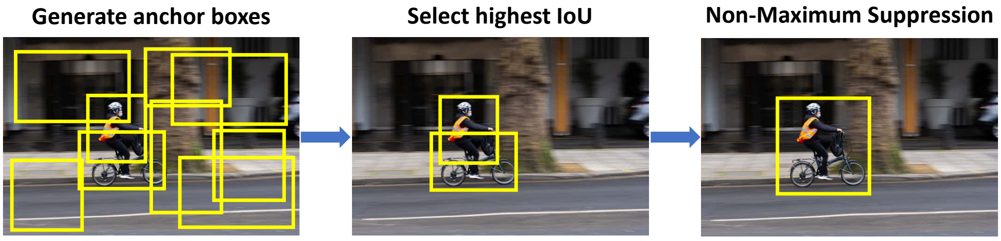


Figure 2.3: The object detection using R-CNN and Fast-RCNN.

sification performance.

Nevertheless, transformers still exhibit similar characteristics to conventional methods in that they detect objects located in the middle, foreground, or large within a scene. This limitation can lead to difficulty identifying smaller or background objects, which is often crucial in disaster detection scenarios.

Object detection methods (Girshick et al., 2014; Girshick, 2015) use the state-of-the-art techniques to generate anchor boxes, as illustrated in Figure 2.3, before selecting the ones with the highest Intersection over Union (IoU) scores. After this selection, the final anchor boxes are refined using the Non-Maximum Suppression technique to fit the objects, resulting in precise detection boundaries optimally.

Therefore, the state-of-the-art method for generating anchor boxes has inspired the author to use masked regions in the image to shift attention in this research. By applying the concept of anchor boxes to masked regions, the aim is to direct the model's attention more effectively towards areas of interest, improving the accuracy of disaster detection.

Chapter 3

Image Captioning Features in Attention

3.1 Introduction

Conventional computer vision techniques detect objects in images by analyzing surrounding pixels to extract information such as texture, intensity, and color (Ofli et al., 2021; Li et al., 2022; Di Biase et al., 2021). However, relying solely on these simple image features can be quite challenging when the object in question is related to a disaster. Disasters involve misplaced natural objects, making them difficult to distinguish using traditional pixel-based methods. In contrast, text tokens generated from image captioning (Johnson et al., 2016) can provide more detailed information about the actions and context of the related objects in the scene, enhancing the detection capabilities.

Computer vision has undergone significant advancements over the years, with Convolutional Neural Networks (CNN) emerging as the most widely used architecture for addressing various challenges in the field (Vaswani et al., 2017). Numerous techniques have been developed for different purposes, such as semantic segmentation, object detection, image classification, and anomaly detection. These include U-Net (Meena et al., 2022; Soares et al., 2020; Liu et al., 2020), ResNet50 (Ofli et al., 2021, 2022), VGG16 (Li et al., 2022), and others (Can et al., 2019). Anomaly detection has found applications in various domains, including the automotive industry and inspection tasks. For example, rescue robots employ sensors or monocular cameras to detect obstacles (Akamine et al., 2022)), enabling them to calculate efficient and safe paths for navigation. The majority of anomaly detection research focuses on identifying unseen or abnormal objects using Generative Adversarial Networks (GAN) for image re-synthesis. These synthesized images are then compared with semantic maps, which are typically generated using CNN-based segmentation techniques (Lis et al., 2019; Kamoi et al., 2021; Ohgushi et al., 2020; Di Biase et al., 2021; Chan et al., 2021).

In recent years, the computer vision field has increasingly adopted Natural Language Processing (NLP) techniques, particularly the attention mask

(Vaswani et al., 2017), to address various challenges. The performance of these NLP-inspired approaches has surpassed that of traditional neural networks. Since 2021, the attention mask or transformer models have gained significant popularity in computer vision challenges. However, existing literature on predicting disaster images has primarily focused on pixel-based features for image classification, as represented in Figure 3.1 (Ofli et al., 2022, 2021; Pennington et al., 2022; Tanatipuknon et al., 2021).

Abnormal scenes are unusual situations that deviate from our daily experiences. These atypical occurrences can have detrimental effects on our regular activities. One such example is landslides, which commonly take place in areas with steep slopes. Landslides can cause significant disruptions to transportation and infrastructure, such as buildings, highways, and roads. In the context of abnormal detection, particularly in the case of landslides, identifying the affected regions is crucial for decision-makers. By analyzing image data captured by unmanned aerial vehicles (UAVs), they can assess the extent of the damage and develop appropriate plans for recovery and reconstruction in the impacted areas.

In addition to the development of anomaly detection techniques, researchers have also focused on comparing the generated images from various methods with re-synthesized, semantic, and original RGB images to identify anomalous objects appearing in the scene (Lis et al., 2019; Kamoi et al., 2021; Ohgushi et al., 2020; Di Biase et al., 2021; Chan et al., 2021). However, anomaly detection in disaster scenarios, particularly in the case of landslides or mudslides, presents significant challenges and complexities. Landslides occur when soil loses its stability on steep slopes (Nefeslioglu et al., 2008) and subsequently falls down the mountainous area. This phenomenon poses difficulties for models attempting to segment or classify the image accurately. In such cases, the model may incorrectly identify the scene as ordinary or misclassify the soil and forest as non-anomalous regions instead of recognizing the presence of a landslide or mudslide. For instance, when analyzing an anomalous image of a landslide where soil and trees have slid onto a road, the segmentation process may classify the landslide region as trees or forest. Conversely, in a normal image, the model might correctly define the objects as soil or water. This discrepancy highlights the challenges associated with accurately detecting and classifying anomalies in complex disaster scenarios.

Classifying images is a traditional problem that the author has faced and solved quite well. For instance, the author have successfully classified various types of leaves (Sardogan et al., 2018) and animal classification (Trnovszky et al., 2017), the classification of natural disaster-related prob-

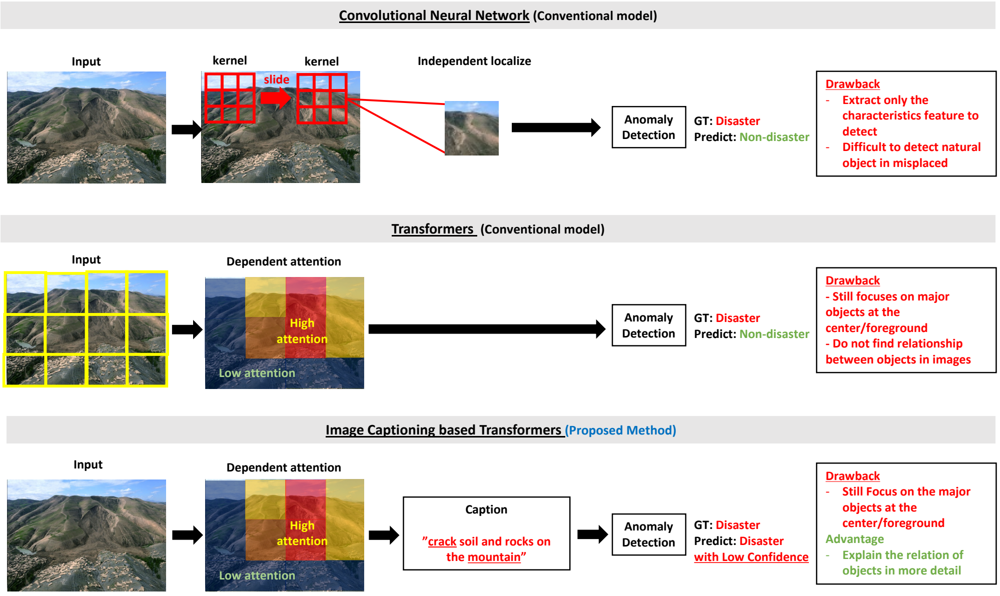

Convolutional Neural Network (Conventional model)
Input
kernel
slide
kernel
Independent localize
Anomaly Detection
GT: Disaster
Predict: Non-disaster
Drawback
- Extract only the characteristics feature to detect
- Difficult to detect natural object in misplaced
Transformers (Conventional model)
Input
Dependent attention
High attention
Low attention
Anomaly Detection
GT: Disaster
Predict: Non-disaster
Drawback
- Still focuses on major objects at the center/foreground
- Do not find relationship between objects in images
Image Captioning based Transformers (Proposed Method)
Input
Dependent attention
High attention
Low attention
Caption
"crack soil and rocks on the mountain"
Anomaly Detection
GT: Disaster
Predict: Disaster with Low Confidence
Drawback
- Still Focus on the major objects at the center/foreground
Advantage
- Explain the relation of objects in more detail

Figure 3.1: The different between Conventional CNNs, Transformers in ViT, and Image Captioning Features Transformers.

lems, such as flooding, landslides, and mudslides, remains a significant challenge. Researchers often utilize Unmanned Aerial Vehicles (UAVs) to investigate these images for evaluating and responding to rescue victims or to recover damaged areas. However, the precision and accuracy of these methods are not yet satisfactory (Ofli et al., 2022, 2021; Pennington et al., 2022), due to the images being derived from natural objects that consist of familiar elements encountered daily, such as soil, rocks, trees, water, or rivers. Researchers attempt to classify these situations by training models on ordinary and irregular datasets (Ofli et al., 2022, 2021; Pennington et al., 2022) using computer vision techniques, but they still struggle to achieve effective classification.

The integration of CNN architectures and GAN-based approaches has revolutionized anomaly detection, allowing for more accurate and efficient identification of irregular or unexpected elements in various scenarios. By leveraging the power of deep learning and generative models, researchers and practitioners can develop robust systems capable of detecting anomalies in real-time, thereby enhancing safety, quality control, and decision-making processes across a wide range of applications.

To address the challenges of accurately classifying complex or ambiguous images, the author proposes a novel method that combines image-to-text techniques with classification algorithms. By generating text descriptions of images, this approach aims to achieve more precise and reliable image classification results. The generated text descriptions can provide valuable insights into the underlying image features that contribute to the classification decision, making the process more intuitive and interpretable. The proposed method offers several advantages over traditional pixel-based classification techniques. By utilizing image-to-text methods, more meaningful features can be extracted from the images, leading to more robust and accurate classification results. Additionally, the generated human-readable text descriptions enhance the interpretability of the classification process, providing researchers with a better understanding of the factors influencing the classification decision.

Furthermore, our proposed method has several advantages over traditional pixel-based classification techniques. The author can extract more meaningful features from the images using image-to-text methods, leading to more robust and accurate classification results. Moreover, our method is more intuitive, as it generates human-readable text descriptions that can provide insights into the underlying image features that contribute to the classification decision.

Furthermore, our proposed method has several advantages over traditional pixel-based classification techniques. The author can extract more meaningful features from the images using image-to-text methods, leading to more robust and accurate classification results. Moreover, our method is more intuitive, as it generates human-readable text descriptions that can provide insights into the underlying image features that contribute to the classification decision.

The author first constructed a dataset from the British Geological Survey (Ofli et al., 2022, 2021; Pennington et al., 2022), which provided us with the landslide images dataset. The author then labels the dataset with text caption in each image as input for the image captioning model. After that, I will use the text caption for the text classification model. Moreover, this technique is state-of-the-art that use language which could explain more detail than pixel-based to make machines understand the surrounding situation and classification anomaly images. Our main contribution is threefold:

- The author creates the image captioning dataset. The dataset includes image captions for the image captioning model and the images from YouTube and Google combined with the British Geological Survey dataset.
- The author proposes to use the image captioning model to generate the text caption to explain the detail of the images instead of using the traditional models to classify the landslide images.
- The author presents the performance of our framework over this new dataset. Furthermore, I then compare it with the traditional method

such as ResNet50 (Ofli et al., 2021) and Vision Transformer (ViT) (Dosovitskiy et al., 2020). This will evaluate our framework, Convolutional Neural Networks (ResNet50), and Transformers model, which are different frameworks.

3.2 Related work

The anomaly prediction challenge is a challenge in the computer vision field. Recently, most research in anomaly prediction mainly focuses on anomaly object detection, which uses for inspection. Anomaly detection uses generated images to compare with the original images to find strange things. In this research, I first focus on anomaly classification of landslide disaster images. Therefore, I aim to study anomaly prediction, image-to-text, and image classification.

3.2.1 Anomaly Prediction

Most research in anomaly detection aims to find irregular objects or defects on the target stuff. In many related works, they tried to establish the original images without strange objects or defects and compared the generated images with the original images (Lis et al., 2019). In the preliminary proposal shows the anomaly detection using the RGB images into the CNN for semantics images. Moreover, the result from CNN would be an input for the Generative Adversarial Networks (GAN) to generate the RGB images in which there are no anomaly objects or flaws on the target objects. They assumed that the semantic maps could not detect abnormal objects well in this situation. Lastly, discrepancy networks would compare the original, semantic, and re-synthesized images to find the different areas that will be defined as anomaly objects. For this reason, the anomaly objects should be smaller than the circumstance objects in the images, which could make the semantic maps quite clear and without the strange objects. Otherwise, the re-synthesized images would generate the RGB images with the irregular objects, and I then could not discriminate between generated images and original images.

After the anomaly detection using generated images to compare the discrimination (Kamoi et al., 2021; Ohgushi et al., 2020; Di Biase et al., 2021; Chan et al., 2021), they used the various layers to extract features in the images to discriminate with the original image. There are many feature extraction techniques, such as softmax entropy, softmax distance, and perceptual difference; they are then compared with the re-synthesized or original images to specify the irregular objects. Anomaly detection is also a robust challenge in the medical field, tried to find the abnormal tissues from grayscale images. This research uses GAN to train only standard image dataset mapping with a Z-uniformly vector (Schlegl et al., 2017). This mentioned study aims to generate the image to be the typical image. Using the discriminating function, they then use the generated images to compare with the original images, which are defined as the anomaly area for irregular tissues. Most anomaly research still aims to generate the usual situation of images before comparing the generated images with the original images to specify anomaly regions.

3.2.2 Image to Text

Image captioning, also known as image-to-text or visual captioning, is the task of generating a textual description for a given image. This area of research has gained significant attention in recent years due to its potential applications in various fields, including computer vision, natural language processing, and robotics. One of the seminal works in image captioning is the "neural image caption" (Xu et al., 2015), which employed a convolutional neural network (CNN) to extract visual features from images and a long short-term memory (LSTM) network to generate textual descriptions. The model was trained on the Microsoft Common Objects in Context (COCO) dataset and achieved impressive results in terms of BLEU-4 scores, which measure the similarity between the generated captions and the ground truth captions.

Since the introduction of the neural image caption, numerous studies have been conducted to further improve the performance of image captioning models. For instance, Zhu et al. (Zhu et al., 2018) proposed a self-attention mechanism for image captioning, which allowed the model to selectively focus on different image regions while generating captions. This approach enhanced the model's ability to capture and describe the most relevant aspects of the image.

Furthermore, researchers have explored the use of pre-trained language models in image captioning. Torrey and Shavlik (Torrey and Shavlik, 2010) utilized a transformer-based language model that was pre-trained on large-scale text and image data to generate captions. The results demonstrated that the model outperformed previous state-of-the-art models on various image captioning benchmarks, highlighting the effectiveness of leveraging pre-trained language models in this task.

In conclusion, image captioning is a rapidly evolving field of research, and deep learning techniques have revolutionized the way textual descriptions are generated for images. While significant progress has been made, generating detailed and specific captions remains a challenging task. However, by leveraging advanced techniques such as transformer-based models, researchers are working towards developing more sophisticated and accurate image captioning systems. The better the generated caption texts are, the better the results of image classification will be, as the captions provide valuable semantic information that can aid in the classification process.

3.2.3 Image Classification

Image classification, a fundamental problem in computer vision, involves assigning predefined labels to images. This task has numerous applications in various fields, including medicine, surveillance, and autonomous driving. The most advanced models currently employ deep convolutional neural networks (CNNs) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Dhruv and Naskar, 2020) with multiple layers to extract hierarchical features from images, as illustrated in Figure 3.1 in the above part. These models have achieved significant improvements in image classification performance. Recently, researchers have investigated the use of attention mechanisms (Vaswani et al., 2017; Dosovitskiy et al., 2020) to further enhance image classification results. Additionally, transfer learning techniques have been explored to improve the performance of image classification models when the amount of labeled training data is limited. A common approach is to use pre-trained models, such as those trained on the ImageNet dataset, and fine-tune them for specific tasks (Simonyan and Zisserman, 2014).

While deep learning techniques have revolutionized the approach to solving image classification problems, classifying features in natural environment images, particularly in disaster cases such as landslides and flooding, remains a challenging task. This is especially true for bird's-eye-view images captured by drones, as shown in the proposed framework in Figure 3.1 in the below part. These images often contain similar objects within the scene, making it difficult to accurately distinguish and classify them.

3.3 Methodology

In this section, the author would like to explain this in three parts. First, the dataset I used in this study came from the British Geological Survey (BGS), and the image data was extracted from the frame in a YouTube video. Next, I explain the image-to-text or image captioning model; The author used Vision

Table 3.1: The statistics of our dataset.

Data source	Training	Validation	Testing	Total	Type
YouTube	44	12	13	69	Anomaly
BGS	1,690	200	211	2,101	Anomaly
Sum	1,761	205	215	**2,170**	Anomaly
Kaggle	3,514	400	405	4,319	Normal
Sum	3,514	400	405	**4,319**	Normal
Total	5,280	605	620	**6,489**	Both

Encoder Decoder (VED) model-based transformer model. Then, I classified the image using text explanation instead of image features classification which part two and three are represented in Figure 3.2.

3.3.1 Dataset

In this study, I will utilize an image dataset and label each image with a text caption. The image dataset is sourced from three different origins, as shown in Table 3.1. Firstly, I collected images from YouTube videos related to landslides and extracted frames from the video data. As a result, the YouTube dataset consists solely of anomaly images before labeling them with text captions in the subsequent stage. The second data source is the British Geological Survey (BGS) (bgs, 2023), which provided landslide images that I then labeled with text captions. Lastly, the common scene image dataset was obtained from Kaggle, which provided a dataset of 4,319 images, as mentioned in Table. 3.1. It is important to note that the datasets from YouTube and the British Geological Survey (BGS) contain landslide or abnormal images, while the Kaggle dataset consists of regular scene images.

These three data sources were employed to train, validate, and test the models in our framework. As shown in Table 3.1, the anomaly images from YouTube and BGS comprise 2,170 images, while the typical images from Kaggle amount to 4,319 images. The author labeled each image in this dataset with a text caption, using common words such as trees, soil, rocks, water, river, or lake. Furthermore, I utilized these object words to describe the surrounding objects and their positions, as this approach would facilitate the model's understanding of the scene's context.

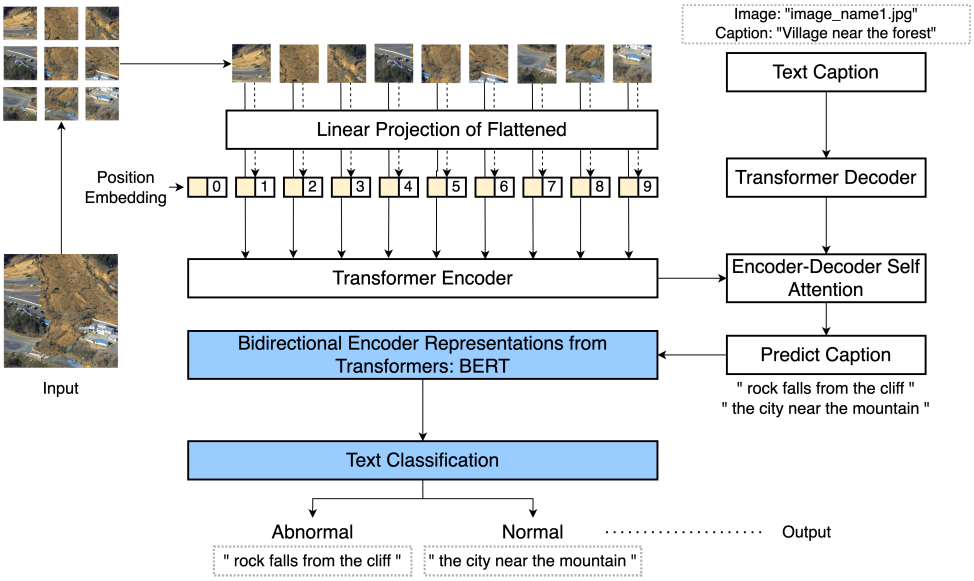


Image: "image_name1.jpg"
Caption: "Village near the forest"
Text Caption
Linear Projection of Flattened
Position Embedding
0
1
2
3
4
5
6
7
8
9
Transformer Decoder
Transformer Encoder
Encoder-Decoder Self Attention
Input
Bidirectional Encoder Representations from Transformers: BERT
Predict Caption
" rock falls from the cliff "
" the city near the mountain "
Text Classification
Abnormal
Normal
Output
" rock falls from the cliff "
" the city near the mountain "

Figure 3.2: Our network architecture.

3.3.2 Image Captioning

Image captioning is achieved by using a Transformer model within a Vision Encoder-Decoder (VED) (Dosovitskiy et al., 2020; Rothe et al., 2020; Li et al., 2021) framework, which integrates computer vision and natural language processing methodologies. Vision Encoder-Decoder (VED) uses image features to generate explainable features, such as text captions, before the classification stage. This approach leads to more meaningful features for classifying natural objects. This approach generates a textual depiction of an image. The framework consists of two primary components: a visual encoder and a textual decoder. The encoder receives an image as input and transforms it into a collection of feature vectors, which are subsequently transmitted to the decoder. The decoder utilizes these vectors to produce a sequence of words by employing a word embedding that includes the word's position, resulting in a sentence that describes the image as a whole. The transformer model is employed in the decoder to produce the textual depiction. The process entails instructing the model using a substantial dataset consisting of pairs of images and their accompanying textual descriptions, with the aim of comprehending the connections between the visual characteristics of the images and the related written descriptions. (Zhou et al., 2020).

During the training process, The datasets for Disaster and Non-disaster were balanced because the number of images in each dataset was quite different. then the model acquires the ability to generate a written caption that is closely related to the image. This model utilizes an attention mechanism, enabling it to concentrate on particular portions of a picture when creating each word. The author applied transfer learning by utilizing a pre-trained Vision Transformer (ViT) model (Dosovitskiy et al., 2020) to fine-tune the encoder part, while the decoder part of the data set was enhanced using Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). The two components are depicted in the upper section of Figure 3.2.

3.3.3 Text Classification

In the lower section of Figure 3.2, the process of text classification using a Transformer model in Bidirectional Encoder Representations from Transformers (BERT) consists of using a pre-trained Transformer model to encode the text input, followed by fine-tuning the model for a particular classification task. Initially, the BERT model performs pre-training on a large dataset of text using an unsupervised learning methodology. During the pre-training phase, the model learns the ability to anticipate the missing words in a sentence and determine the relationship between two sentences. Pre-training allows the model to convert text input into a comprehensive contextualized representation that captures the meaning and connections between words in the text (Zhou et al., 2020).

The BERT model, which was previously trained, is adjusted to perform a particular classification task by using a small dataset with labeled examples for text categorization. In this case, I classify the caption text into two distinct types: normal and abnormal. The final layer of the model is substituted with an output layer. Subsequently, the complete model is trained using the labeled data. During the training process, the model's parameters are adjusted in order to minimize the loss. The author employed text captioning to refine the BERT pre-trained model specifically for common scene caption text and landslide caption text.

The fine-tuned BERT model encodes incoming text caption inputs into a sequence of contextualized representations. It then uses a task-specific output layer to make predictions and classify the input. The use of a Transformer model in BERT for text classification has achieved exceptional performance on diverse benchmark datasets by leveraging on the pre-trained model's capacity to encode contextualized text representations. Thus, I determine the text caption from the VED model discussed before in order to classify scenes or images as either standard or abnormal.

3.4 Experiments

3.4.1 Settings

The author divided the dataset into 80%, 10%, and 10% for training, validation, and test splits, respectively. The author utilized the Transformers library from Huggingface to construct our model. The VED model utilized a pre-trained ViT model as an encoder and BERT as a decoder component. In addition, I utilized the ViT pre-trained model as the feature extractor. The loss function used was binary cross-entropy, which was applied to the score. The vocabulary size was 50, 256 and the batch size was 4. The author utilized a pre-trained BERT model with an Adam optimizer for text categorization. The learning rate and batch size were configured to 10 e-5 and 16, respectively. Ultimately, I labeled the VED section using visual representations and textual explanations that detail the locations of the nearby objects. Subsequently, I classified the text captioning in the concluding section.

In addition, I set up the image classification model (ResNet50) with a learning rate of 10e-4 and a weight decay of 10e-3, as stated in the source (Ofli et al., 2021). The ResNet50 model previously served as the standard for evaluating landslide classification issues. The author calculated the F1-score, as well as the accuracy, precision, and recall. The comparatively high recall rates suggest that the performance for detecting a landslide or massive object in the image could be exceptional.

3.5 Result and Discussion

3.5.1 Image Captioning Prediction

The image captioning model the author trained by fine-tuning a BERT model for decoding and a ViT model for encoding was able to accurately predict the text caption. In this research, a limited number of tokens were used to predict 215 images from the BGS and YouTube datasets, out of a total of 405 images in the regular scene testing set from Kaggle and the landslide scene. During the training of the model, I set the token prediction limit to be the same as the maximum token. As a result, this text caption is able to accurately depict the position of objects in the given environment.

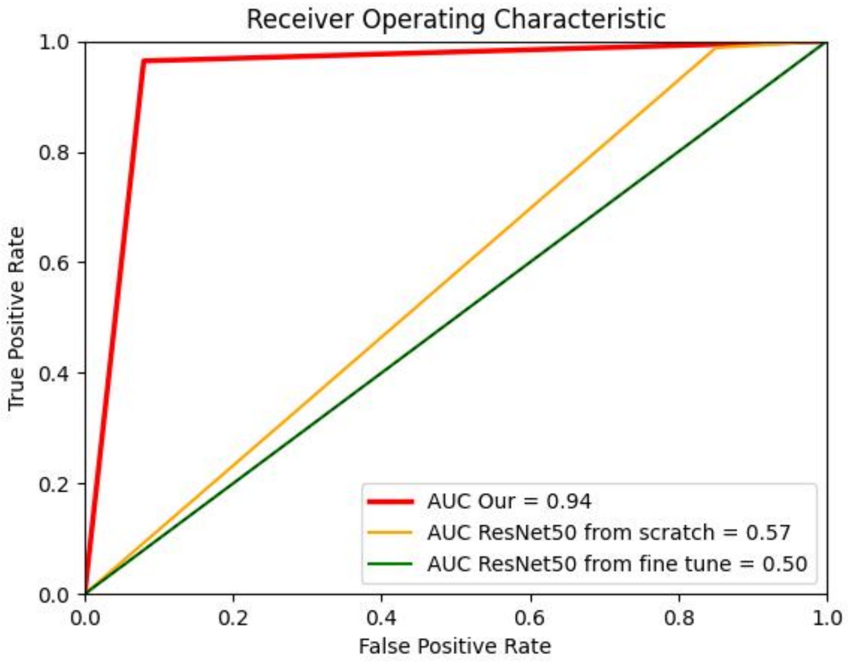


Figure 3.3: AUC between ResNet50 (Ofli et al., 2021) and our proposed framework.

3.5.2 Classification Prediction

The classification issue of landslides is complex due to the similarity between items in normal images and landslide phenomena. Previous studies employed convolutional neural networks to address this issue. For this investigation, I employed ResNet50, a convolutional neural network renowned for its ability to classify landslide photographs. In this result consists of the proposed method using VED from image to caption, ResNet50 training from scratch, and ResNet50 fine-tuning from ImageNet pretrained model. According to Table 3.1, this experiment uses the testing dataset for Disaster (landslide) images from YouTube and the British Geological Survey (BGS), and for Non-disaster (non-landslide) images from Kaggle. As depicted in Figure 3.3, the Receiver Operating Characteristic (ROC) curve of ResNet50 exhibits a 45-degree angle at the center line, while the proposed method is positioned above the middle line. In addition, the ResNet50 model has an Area Under the ROC Curve (AUC) of 0.50, indicating that it is unable to differentiate between a normal image and an image depicting a landslide in the context of landslide classification. However, our suggested model has an Area Under the Curve (AUC) value of 0.94. This indicates that the model is able to accurately differentiate

Table 3.2: The performance for ResNet50 (Ofli et al., 2021) and our proposed framework.

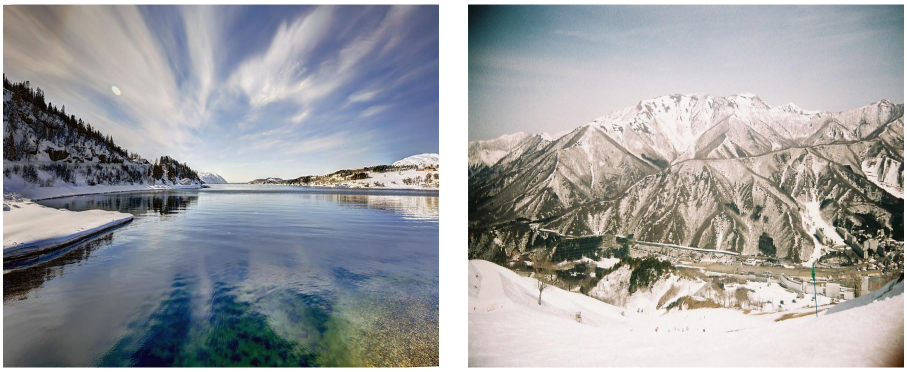
Method	Accuracy	Precision	Recall	F1-Score
Fine-tune (Ofli et al., 2021)	67.58	67.58	**100.00**	80.65
Scratch (Ofli et al., 2021)	71.61	**98.81**	70.77	82.47
Our method	**95.00**	96.19	96.42	**96.31**

between normal images and landslide images with a 94% success rate. Hence, I may infer that our new approach outperforms the existing methods in the task of classifying landslides.

In addition, I developed the confusion matrix to determine the accuracy, precision, recall, and F1-score for evaluating the performance of each approach. The author used the normal and landslide datasets to do fine-tuning on the ResNet50 model in ResNet50. The author modified the ResNet50 model by combining a dense layer consisting of 64 units, utilizing the pre-trained ImageNet model. The author have fine-tuned the final two dense layers to classify photos into two categories: common images and abnormal images. The outcome shows an accuracy rate of 67.58%, indicating that ResNet50 was able to accurately predict just 67.58% of the data in the table refer to table 3.2. Regarding other indices, the precision is 67.57% and the recall is 100.00%. The precision refers to the accuracy of the model in distinguishing true positives (TP) from false positives (FP), with a ratio of 67.58%. The recall refers to the adjustment of the model's true positive (TP) to false negative (FN) ratio, which is 100.00%. The F1-score is the mean value of the single metric that combines precision and recall, which is 80.65%. Nevertheless, I must mention that the ResNet50 model, which I used for landslide classification, is unable to differentiate between two classes with an Area Under the Curve (AUC) support. This is because I just fine-tuned the last two layers, and the ImageNet pre-trained model is not adequate for landslide classification. Alternatively, I attempted to compare the performance of the ResNet50 model trained from scratch, which showed superior results as indicated in the table. The results may be found in Table 3.2, with an Area Under the Curve (AUC) value of 0.57.

Furthermore, our proposed technique provides confusion matrix indices with accuracy, precision, recall, and F1-score values of 95.00%, 96.19%, 96.42%, and 96.31%, respectively. As a result, our approach demonstrated improved performance in classifying familiar and landslide scenarios. Therefore, our proposed method utilizes computer vision to convert images into text, allowing us to determine the position of objects mentioned in the text caption. Additionally, we utilize sentiment analysis to predict the classification of landslides, which can be difficult due to the similarities between landslide regions and objects such as land or trees.

The result in Figure 3.4 (left) and Figure 3.4 (right) show that the caption from VED explains the correct aspect, making the classification a normal case.

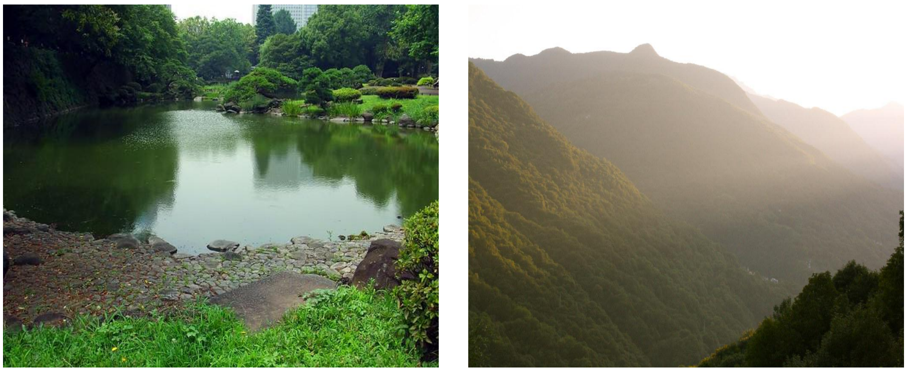


clear river with snowy hill alongside

snowy mountain behind the tree

Figure 3.4: The result in normal class, and the model predicted to normal.

Nevertheless, our proposed solution, which utilized an image captioning technique, may produce incorrect text captions in both false negatives and false positives. For example, Figure 3.5 (left) illustrates that the text of the caption states "soil and rocks fall to the pond", yet the actual image displays the presence of the pond along with soil, rocks, and trees in its surrounding. Similarly, Figure 3.5 (rifgt) displays the outcome of text captioning as "soil slide down from the hill", but the actual image shows a mountainous region with trees.

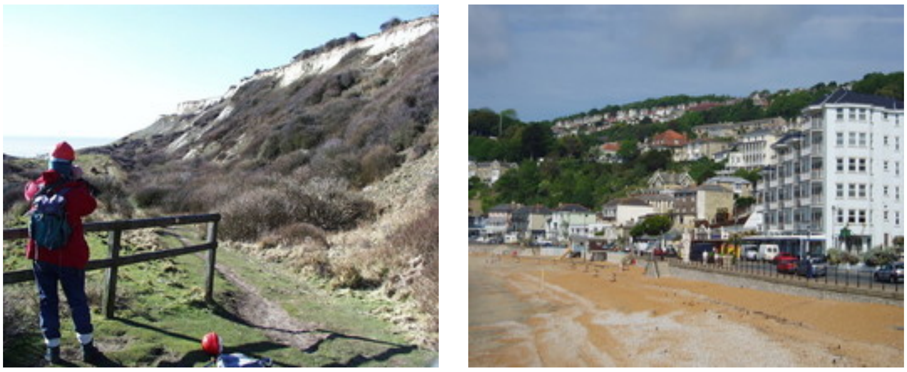


soil and rocks fall down to the pond

soil slide down from the hill

Figure 3.5: The result in normal class, but the model predicted to abnormal.

For false positive scenarios, Figure 3.6 (left) indicates that a guy witnesses the landslide at the high ground in the mountain region, but the model could only inform us that ”man on the hill." Thus, it can be inferred that the model is capable of capturing only the primary items. Furthermore, the model attempts to construct the primary caption due to the constraint of prediction tokens, which results in the model not producing a detailed textual phrase. In contrast, in Figure 3.6 (right), the amount of soil from the landslide on the road is rather minimal.

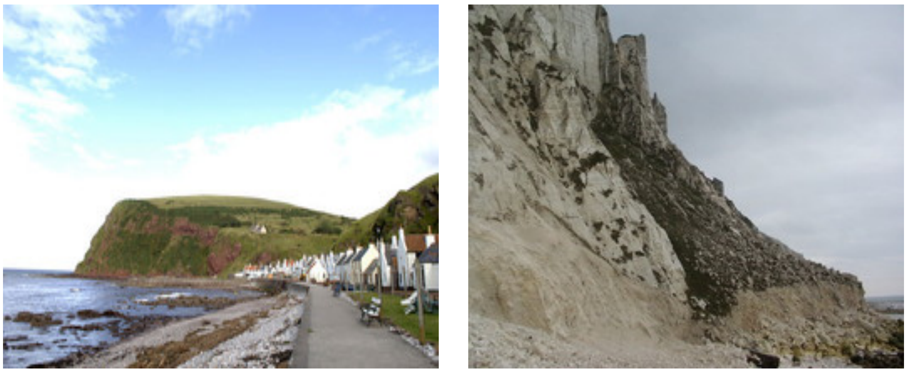


man on the hill

building along the road

Figure 3.6: The result in abnormal class, but the model predicted to normal.

As a result, the model primarily emphasizes the larger items rather than the smaller ones when generating the text caption. To enhance the model, the author can optimize its performance by expanding the prediction token limit. Furthermore, it is imperative to include a more comprehensive and specific description in the caption label in order to enhance performance in this particular scenario. The figures representing the true positive are displayed in Figure 3.4 (left) and 3.4 (right), while the figures representing the true negative may be seen in Figure 3.7 (left) and 3.7 (right).



damaged hill caused by soil slide

soil slide down from the hill

Figure 3.7: The result in abnormal class, and the model predicted to abnormal.

3.6 Conclusion

The author have presented a new method for image classification, specifically in the context of landslides. The challenge involves accurately differentiating between landslides and natural objects like dirt, rocks, and trees, which poses a significant difficulty in classification. The author utilized the vision encoder-decoder approach for generating image captions. This approach utilized pre-trained Vision Transformers (ViT) as the encoder and Bidirectional Encoder Representations from Transformers (BERT) as the decoder. Subsequently, the author employed the text caption to categorize the feeling of the visual scene. Furthermore, the author applied the ResNet50 model for landslide picture classification, which had been previously utilized as a benchmark in other studies.

The AUC analysis shows that ResNet50 is unable to differentiate between normal and landslide photos. Our method achieved a 94% accuracy in distinguishing between normal and landslide photos. In addition, the ResNet50 model, after fine-tuning with ImageNet, achieved an accuracy of 67.58%, precision of 67.58%, recall of 100.00%, and F1-score of 80.65%. The ResNet50 model, trained without using pre-existing weights, achieves an accuracy of 71.61%, precision of 98.81%, recall of 70.77%, and F1-score of 82.47%. Our suggested model achieves accuracy, precision, recall, and F1-score of 95.00%, 96.19%, 96.42%, and 96.31%, respectively. These performance metrics surpass those of ResNet50. However, the proposed method is unable to categorize photos that contain objects with complex locations. ResNet50, which the author trained from scratch, is more effective at classifying landslide photographs compared to ResNet50 fine-tuned using ImageNet.

In addition, the model mainly highlights the key objects to clarify the surrounding circumstances rather than the small details. For instance, the view includes trees, rocks, and dirt surrounding the pond. However, the captioning text refers to the entire scene as falling down to the pond, as shown in Figure 3.5. Furthermore, it is worth noting that while the landslide takes place at the highest point of the mountain region, the accompanying image in Figure 3.6 depicts a guy standing atop a hill. This outcome illustrates that the model primarily prioritizes the principal things in the image as the central figures of the scene, rather than emphasizing the smaller objects and intricate details of the image. Based on the outcome, employing picture captioning for the categorization of intricate sceneries yields superior performance and enhanced efficiency compared to the pixel-based approach. Furthermore, I might enhance the performance by augmenting the constraint on text captioning tokens and providing more comprehensive labeling for captioning the scenes.

3.7 Contribution

The contributions of the work presented in this chapter were:

- Creating the image-caption dataset for Vision Encoder Decoder image captioning model.
- Extract features from Image caption from Vision Encoder Decoder (VED) to capture the action features of natural scenes helps to distinguish between non-disaster and disaster scenes.
- Classifying disaster and non-disaster scenes based on the action features obtained from the Vision Encoder Decoder (VED) model.

Chapter 4

Shifted Attention

4.1 Introduction

Traditional image classification tasks usually require determining only one category for each image, as mentioned in several research papers (Sardogan et al., 2018; Trnovszky et al., 2017). Several methods in this field have utilized neural networks to extract image characteristics, thereby aiding the classification process (Meena et al., 2022; Soares et al., 2020; Liu et al., 2020; Ofli et al., 2021, 2022; Li et al., 2022; Can et al., 2019). Following the progress in image classification, object detection has become a crucial technique that aims to find more precise classes inside each region of an image. The use of anchor boxes is utilized to precisely define the bounding boxes for target object classes in object detection, as described in the works of (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015). Notwithstanding these progressions, the task of detecting landslides, floods, and wildfires in nature continues to be difficult. The challenge commonly arises due to the frequent indistinct look of affected areas, such as landslide zones resembling regular soil or flooded areas reflecting the appearance of ponds or lakes (Ibrahim et al., 2021; Hernández et al., 2022), which can result in potential misclassification.

However, the development of image captioning has revealed a more subtle aspect of computer vision. This extends beyond basic classification to encompass the representation of the surrounding context of an image. In the past, image captioning approaches typically required extracting features from images using Convolutional Neural Networks (CNNs) and then generating textual captions using Recurrent Neural Networks (RNNs) (Johnson et al., 2016). In recent times, the introduction of transformers (Vaswani et al., 2017) has brought about a significant transformation in this particular industry. This method converts the image into a sequence of image tokens, while the textual caption is translated into text tokens in a similar manner. The tokens undergo processing using an encoder-decoder attention layer called Vision Transformers (ViT) (Dosovitskiy et al., 2020). This layer helps develop connections between the image and the caption content. While current image captioning models have the ability to generate detailed captions about images, which can

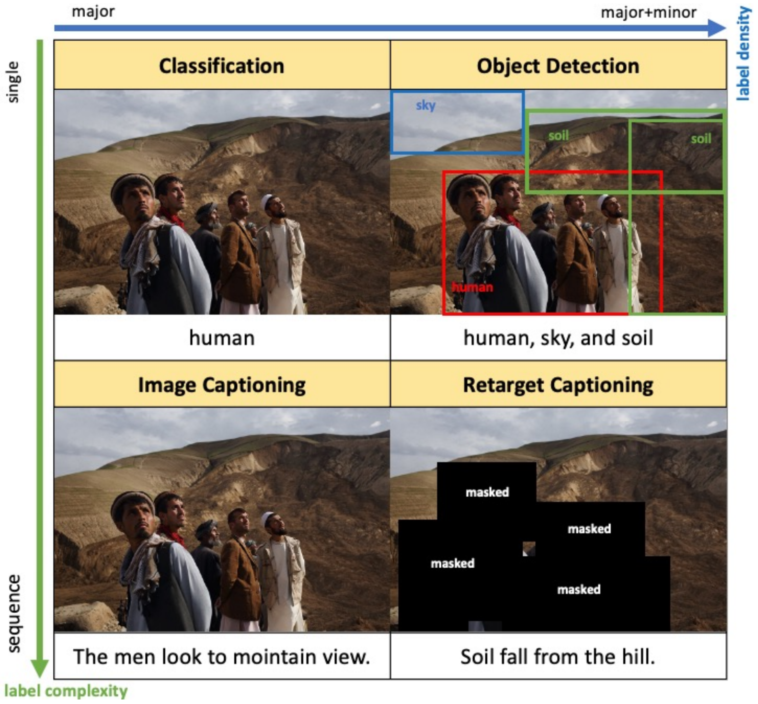


Figure 4.1: Concentrated captioning for more focusing and explanation on the target objective.

improve disaster detection, existing research (Thanyawet et al., 2023) shows that these models mainly concentrate on major objects when generating captions, as shown in Figure 4.1. This approach frequently fails to consider tiny, yet important, areas inside the image. Specifically, aerial photographs that capture large areas may include small locations where disasters are present. Creating a model that can identify both the primary items in a scene and generate descriptions of additional significant elements in the image could facilitate faster emergency response.

Natural calamities such as landslides(Hungr et al., 2014), flooding (Ibrahim et al., 2021; Hernández et al., 2022), or wildfires (Pan et al., 2020) have the potential to transpire in any location and have a profound impact on individuals' lives and day-to-day routines. Landslides, which are most common in hilly areas, are caused by a range of factors such as heavy rainfall, unstable slopes, or seismic activity. Disasters like landslides have the potential to obstruct transit routes and do significant harm to essential infrastructure, including roads, buildings, and highways. Accurate identification of damaged areas is essential for decision-makers to successfully plan for restoration and recovery. Unmanned Aerial Vehicles (UAVs) are crucial in this process, as they provide critical visual data to evaluate damage and inform decision-making. A multitude of computer vision methods, such as detection and classification, have been created to enhance the efficiency and velocity of catastrophe detection, aiming to address these challenges.

This research introduces a retargeting network called RetNet, which guides the model's attention towards specified goal objectives, even if they are little regions of disaster in the photos. Based on the work of (Johnson et al., 2016)Â on dense captioning, the author use anchor boxes and optimize their central positioning, height, and width to determine the first masked areas. In addition, RetNet also has a secondary goal, which is to choose the most suitable anchor boxes for masking that fit with the specific objectives being addressed. This innovative technique allows the model to concentrate on important things, resulting in a more comprehensive and contextual comprehension.

Our contributions in this field are threefold:

- This study presents RetNet, a new approach that refocuses image captioning models' attention. Previously, these models give higher importance to larger and more noticeable things in an image. RetNet, on the other hand, is specifically meant to emphasize less prominent but still important aspects, which is especially advantageous in intricate natural surroundings. Conventional methods can often miss important information in such contexts, but these small features can provide crucial data.

- This study showcases an advanced approach for enhancing anchor boxes, building on the existing Fast-RCNN architecture. This approach specifically emphasizes the precise adjustment of the central position, height, and width of the anchor boxes. Optimizing the detection and labeling of smaller items in real situations is essential, especially as typical picture captioning algorithms may struggle with this task.

- The research expands the validation of the RetNet model to encompass a wide array of disasters, including floods and wildfires. It explores these events from several angles, including airborne, shipborne, and conventional human-captured perspectives. RetNet is utilized to examine regions impacted by calamities, utilizing the model's improved picture captioning and text categorization abilities to effectively detect and categorize different elements and types of disasters. This application showcases the tangible and noteworthy influence of RetNet in realistic situations, specifically in prompt and efficient catastrophe response and evaluation.

4.2 Related works

This section provides an overview of previous research on the development of computer vision techniques for disaster detection. It includes studies on traditional image classification (Sardogan et al., 2018; Trnovszky et al., 2017; Meena et al., 2022; Soares et al., 2020; Ofli et al., 2021, 2022), object detection (Can et al., 2019; Sameen and Pradhan, 2019), object segmentation (Liu et al., 2020; Li et al., 2022), and image captioning generation (Johnson et al., 2016; Castro et al., 2022). Despite the utilization of sophisticated methodologies, accurately discerning indistinct or ambiguous entities in photographs, such as calamities, continues to pose a formidable obstacle.

4.2.1 Advancement in object detection

The utilization of anchor boxes for object detection in images is a state-of-the-art computer vision methodology. This technique prevents the detection of objects with low probability throughout the process of making predictions and generates anchor boxes for every patch in a picture. The approach employs intersection over union (IoU) calculations for each class, as elucidated in seminal publications such as R-CNN, Fast R-CNN, and Faster R-CNN, to retain just the most probable detection (Ren et al., 2015; Girshick et al., 2014; Girshick, 2015). During the training phase, it is crucial to assign labels to the object classes and define the boundary boxes as a fundamental aspect of this technique. Just like in image segmentation, where boundary labels are used to extract features before training, this preparation is essential for accurately training the model to generate anchor boxes.

Our methodology provides uniqueness by repurposing the anchor box production process to construct masked regions inside the image, thus expanding upon the concept of anchor boxes. The model is trained to learn the optimal positions and sizes for these anchor boxes, which are then used to cover the original image. The purpose of this strategy is to redirect the model's attention towards the things in the image that are less accurate and are overlooked more frequently. The author aim to enhance the model's ability to detect and classify ambiguous or less prominent features in the visual data by modifying the model's attentional hierarchy.

4.2.2 Transformer-based image captioning

The emergence of transformer models in recent years has represented a notable progress in the industry. Originally designed for machine translation, these

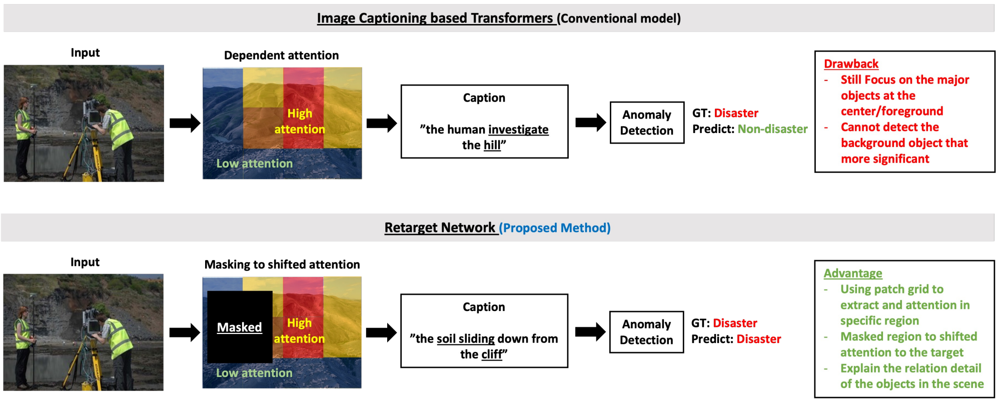


Image Captioning based Transformers (Conventional model)
Input
Dependent attention
High attention
Low attention
Caption
"the human investigate the hill"
Anomaly Detection
GT: Disaster
Predict: Non-disaster
Drawback
- Still Focus on the major objects at the center/foreground
- Cannot detect the background object that more significant
Retarget Network (Proposed Method)
Input
Masking to shifted attention
Masked
High attention
Low attention
Caption
"the soil sliding down from the cliff"
Anomaly Detection
GT: Disaster
Predict: Disaster
Advantage
- Using patch grid to extract and attention in specific region
- Masked region to shifted attention to the target
- Explain the relation detail of the objects in the scene

Figure 4.2: Disaster image classification network architecture between prior works and our proposed method. There are convolution layers and flatten layer for prior work. For proposed method, VED to generate caption token with masked images from Retarget Network for text classification.

models employ attention layers consisting of transformer blocks. These blocks greatly improve the model's ability to concentrate on values that demonstrate important correlations, hence enhancing its concentration. This advancement signifies a significant change in the way visual information is analyzed and understood, providing a more sophisticated and situationally conscious method for computer vision tasks. Transformer models, which have achieved significant advancements in natural language processing, have also been used to computer vision tasks, namely in the area of image classification (Dosovitskiy et al., 2020; Rothe et al., 2020; Li et al., 2021). Utilizing transformer approaches in this particular situation improves the process of extracting features, enabling a more concentrated focus on important elements within images. This advancement signifies a significant progression in the examination of visual data, facilitating more precise and comprehensive understandings.

4.2.3 Challenge in current image captioning

An essential focus of computer vision research involves generating textual descriptions based on input images. RNNs are commonly employed for generating captions, after the use of CNNs for extracting features (Johnson et al., 2016; Castro et al., 2022). This method effectively establishes a coherent storyline for the visual data by linking the textual captions with the retrieved image characteristics. In light of the complexity of the photographs, comprehensive captioning has been employed, particularly due to the frequent inclusion of several objects. By employing anchor boxes, this method divides the image into sections of interest and generates a distinct caption for each zone, resulting in a more comprehensive explanation (Johnson et al., 2016).

The author suggest that an image has the potential to encompass multiple components, each of which warrants a more comprehensive elucidation than a solitary statement. Image captioning offers a more significant method to convey these intricacies, particularly when producing results in terms of human language. This approach is very effective for clarifying uncertain objects, such as those seen in the aftermath of natural calamities. In these instances, it may be necessary to accurately identify or categorize object properties that are solely represented as pixel-based images. Consequently, image captioning can have a crucial role in providing individuals with a more profound comprehension of these intricate scenarios.

4.2.4 Application in disaster management

The 2P2R method(Tantanee et al., 2018), emphasizes that disaster management begins with proactive measures such as constructing or renovating infrastructure to mitigate the impact of disasters. Two examples include constructing sea walls to mitigate tidal impacts (Thomas and Hall, 2015) and enhancing building foundations to better absorb seismic vibrations (Mirzaev et al., 2021; Haseeb et al., 2011). The subsequent stage involves preparation, encompassing pre-event arrangements such as ensuring the availability of food and water provisions and charting evacuation pathways. The rehabilitation phase following an incident involves assessing and rectifying the harm caused, which includes identifying the affected individuals and regions. The response phase involves the issuance of warnings and the evacuation of individuals. The utilization of UAV or helicopter aerial surveillance is quite valuable during this time.

Our research aims to utilize images to identify damaged regions throughout the recovery phase, in order to pinpoint areas that require restoration utilizing our innovative approach. Disasters can cause natural features to be moved to unfamiliar places, which makes it difficult to identify these uncertain things using traditional approaches. Determining whether earth pouring from a cliff and blocking a mountain transportation route qualifies as a damaged area might be challenging. By enhancing the accuracy of item detection, our approach has the capacity to greatly enhance the precision as well as effectiveness of damage evaluation in disaster management.

4.3 Preliminary Investigation

In this section, the author examine how the caption-based approach (Thanyawet et al., 2023)), outperforms pixel-based strategies(Ofli et al., 2021), in terms of classification accuracy. The caption-based approach has the ability to generate more significant characteristics, resulting in improved classification of disaster photographs. Furthermore, the machine-learning models developed by the captioning process (Johnson et al., 2016; Vaswani et al., 2017) have a tendency to prioritize the things that are most prominent in the foreground. This experiment showcases the model's ability to effectively extract important properties from masked and cropped photos, as well as the key attributes of appropriate masking for accurately identifying target items.

4.3.1 Classification (Pixel-based vs Caption-based)

4.3.1.1 Experiment Setup

The objective of this experiment is to illustrate the difficulties faced by classic pixel-based models in classifying disaster images, as they heavily depend on feature extraction for image classification. The author conduct a comparison between pixel-based models, notably ResNet50 (specifically, ResNet50 (Ofli et al., 2021)), and text-based picture captioning features that employ a Vision Encoder-Decoder (VED) framework, in the context of catastrophe scene classification. The author utilized a testing dataset consisting of 620 images, trained the model using 5,280 images, then allocated 605 images for validation. The image dataset comprises a collection of both regular images and landslide scenes.

4.3.1.2 Experimental Results

The results are displayed in Table. 4.1. The utilization of 620 images in the testing dataset provides evidence that the VED technique surpasses traditional models in terms of performance, obtaining an impressive Area Under the Curve (AUC) value of 0.94 (Thanyawet et al., 2023). This is a noteworthy enhancement compared to the performance bar set by ResNet50(Ofli et al., 2021). Nevertheless, the author have seen that conventional ResNet50 models and the VED approach exhibit a tendency to give higher importance to foreground objects, frequently overlooking small catastrophe zones present in the image. This observation prompted us to create a framework specifically aimed at redirecting attention onto the items in the vicinity, such as tiny catastrophe areas, as depicted in Figure 4.2. The Retarget Network aims to optimize the

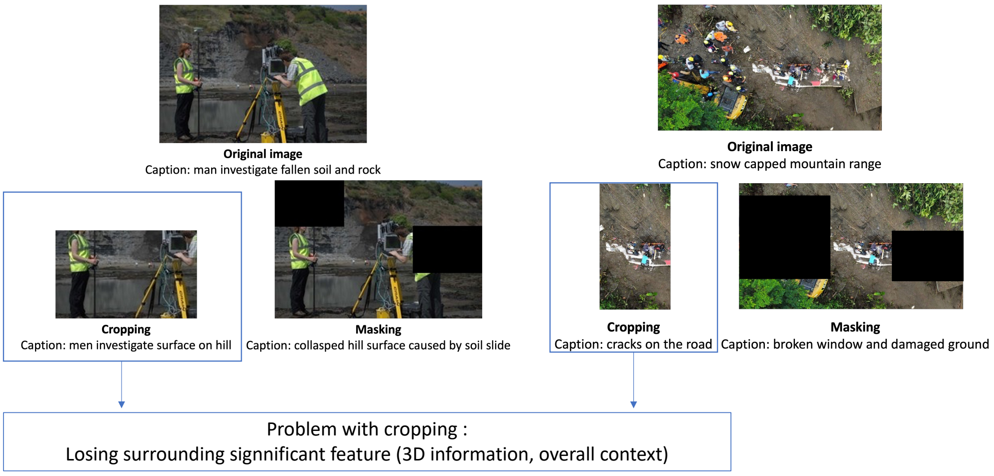

Original image
Caption: man investigate fallen soil and rock
Cropping
Caption: men investigate surface on hill
Masking
Caption: collasped hill surface caused by soil slide
Original image
Caption: snow capped mountain range
Cropping
Caption: cracks on the road
Masking
Caption: broken window and damaged ground
Problem with cropping :
Losing surrounding signnificant feature (3D information, overall context)

Figure 4.3: Landslide image captioning between applying cropping and masking techniques to focus on the affected area in different view aspect.

prioritized regions for enhanced target area detection in catastrophe photos, especially in images of significant scale.

Table 4.1: Comparison of existing methods for image classification

Method	Accuracy	Precision	Recall	F1-Score
ResNet50 fine-tune (Ofli et al., 2021)	67.58	67.58	**100.00**	80.65
ResNet50 from scratch (Ofli et al., 2021)	71.61	**98.81**	70.77	82.47
VED (Thanyawet et al., 2023)	**95.00**	96.19	96.42	**96.31**

4.3.2 Image Captioning from Cropping vs Masking

4.3.2.1 Experiment Setup

The author conducted experiments to compare the effects of cropping and masking, which are two separate techniques used in image editing. Cropping involves trimming the outer areas of an image to focus on specific sections, which can prompt the model to give more attention to often disregarded elements. Masking is the act of concealing specific parts of an image to allow the model to concentrate on the visible elements, which may be less conspicuous yet are crucial. In these studies, the author employ a total of 16 photos that are representative of side view, shipborne, and aerial view images.

4.3.2.2 Experimental Results

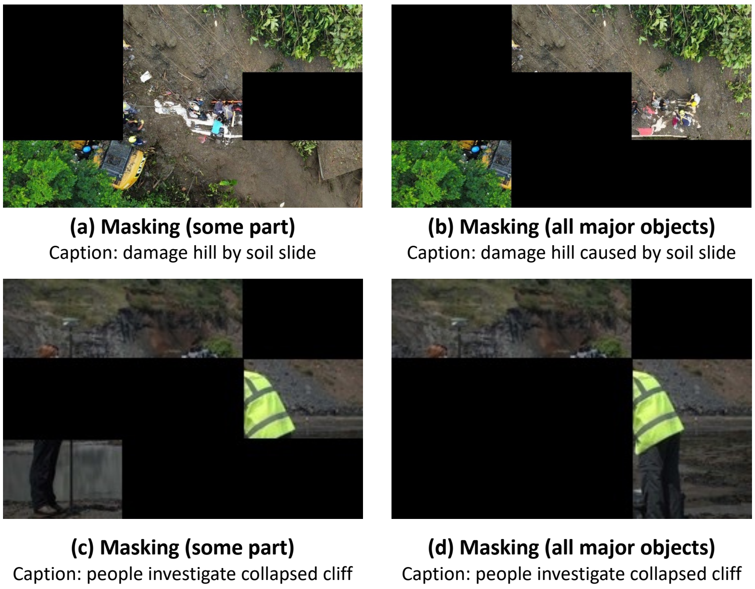
The analysis, depicted in Figure 4.3, demonstrates contrasting outcomes obtained from our two experimental approaches. A notable disadvantage of cropping photos is the loss of essential characteristics from surrounding objects. As a result of this paucity, captions frequently contained false information or lacked enough details to fully explain in everyday language.

Conversely, the masking strategy produced more promising results. The model generated captions that incorporated these secondary things by utilizing masking techniques, which selectively obscured the primary object of interest while keeping other elements visible. This approach facilitated a more thorough comprehension of the scene as the captions elucidated both the single object and its broader context, as well as the interconnections among the different elements. Masking specifically enables the model to concentrate on gaining a more unbiased and comprehensive comprehension of the disaster scenario.

4.3.3 Optimal Masking

4.3.3.1 Experiment Setup

In these studies, the author utilize 16 photos that are representative of side view, shipborne, and aerial view images, similar to the cropping versus masking experiment. Subsequently, the author investigated the impact of selectively concealing specific regions on a model's ability to generate precise image descriptions. By employing a systematic and exhaustive approach, the author analyzed all possible regions to identify the specific attributes that yielded captions that closely matched the actual data. To facilitate the experiment, the author partitioned each image into nine patches and organized them in a 3x3 grid. As a result, 512 distinct masking configurations could be made, or $2^{(3\times3)}$. The author examined the characteristics and trends in the captions created for these masked images in all possible combinations.



(a) Masking (some part)
Caption: damage hill by soil slide

(b) Masking (all major objects)
Caption: damage hill caused by soil slide

(c) Masking (some part)
Caption: people investigate collapsed cliff

(d) Masking (all major objects)
Caption: people investigate collapsed cliff

Figure 4.4: Comparison of images with only some parts masked ((a) and (c)) and with all major objects masked ((b) and (d)).

4.3.3.2 Experimental Results

Our results indicate that achieving precise caption creation does not necessitate the masking of every discernible object in the image. Figure 4.4 depicts an example that demonstrates the similarity between the captions of photographs with partial masking and images with complete masking of the main items. In addition, the author have seen that the process of handling 16 photos using the brute-force masking technique takes over 4 hours of computational time. These findings demonstrate that the use of masking substantially increases the computational requirements, despite its ability to effectively redirect the model's focus. This underscores the necessity for improved methodologies. Therefore, the author present the Retarget Network, an innovative approach aimed at achieving a compromise between computational efficiency and accurate detection.

4.4 Methodology

This section presents the network architecture and pipeline of our proposed RetNet. RetNet is designed to redirect attention from prominent objects to other objects that may be less conspicuous but still play crucial roles in the image. In the training phase, the author balanced the dataset to train for

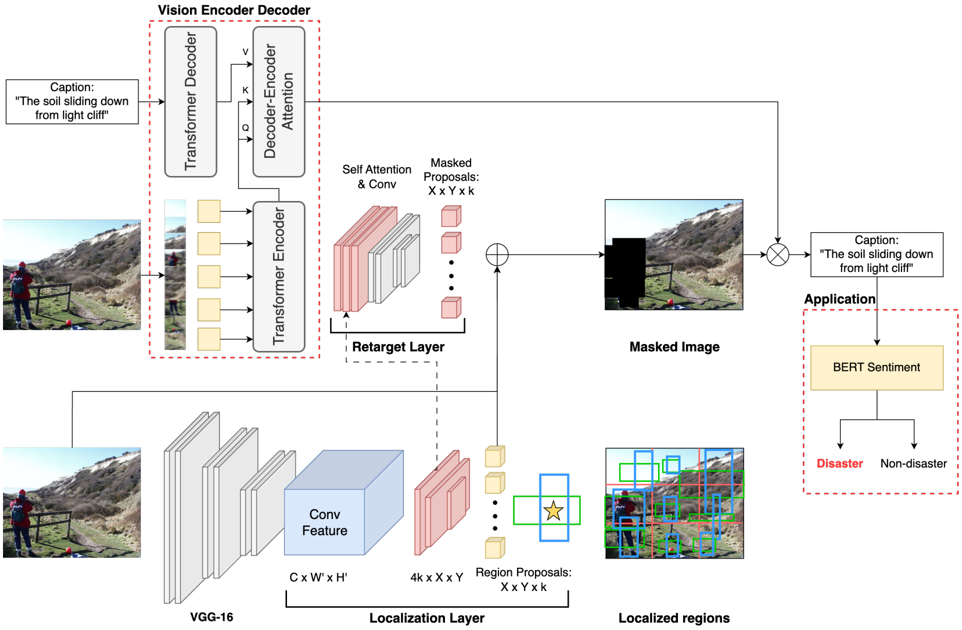


Figure 4.5: Our network architecture.

landslide disaster detection by shuffling the Non-disaster images, which were more numerous than the Disaster images. Then, the RetNet model, which was trained on landslide disasters, was fine-tuned with flood and fire disasters, respectively, before being used for inference to evaluate its performance in the inference phase.

4.4.1 Network Architecture

The architecture of RetNet is depicted in Figure 4.5. Prior to extracting the picture features and converting them into feature maps using the VGG-16 model, the author first train the Visual Encoder-Decoder (VED) for image captioning,(Johnson et al., 2016). In this network, the author introduce two new layers to the RetNet architecture: the Localization layer and the Retarget layer. The retarget layer further enhances the potential masked region candidates generated by the localization layer to get ideal masks. Patches and anchor boxes are employed in the localization layer to generate candidate-masked regions. The image captioning model then redirects its attention by utilizing the retarget layer to identify the most prominent masked regions among the candidates.

4.4.2 Vision Encoder Decoder

Image captioning is performed using a Transformer model within a VED (Dosovitskiy et al., 2020; Rothe et al., 2020; Li et al., 2021) framework generates a caption for an image by combining computer vision and natural language processing techniques. The framework consisted of two primary components: the text decoder and the visual encoder.

- Vision Encoder: The image is split into patches, which are subsequently inputted into an encoder transformers block to generate feature vectors. For this procedure, the author employed a pre-trained Vision Transformer (ViT) model (Dosovitskiy et al., 2020) to encode the images.
- Caption Decoder: The incoming text is processed by the decoder transformer block (Vaswani et al., 2017) to embed text caption labels into feature vectors. During the decoding process, the author employed the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) to incorporate the caption text, along with the use of tokenizers.

The author use a dataset consisting of pairs of images and fixed-length captions to train the network. The objective is to enable the network to understand the connections between the feature vectors of the images and their related textual descriptions (Zhou et al., 2020). The model utilizes learnt parameters during the inference process to generate text captions based on the correlations observed in the picture feature vectors.

4.4.3 Localization Layer

The purpose of this layer is to produce potential areas for concealing regions from the feature map of an image, as described in the study by Long et al. (2015). The author utilize the cutting-edge VGG-16 architecture (Simonyan and Zisserman, 2014; Russakovsky et al., 2015), which consists of 13 layers of 3×3 convolutions alternated with 5 layers of 2×2 max pooling, to extract the feature map I. As a result, an input image with dimensions $3 \times W \times H$ is transformed into a feature map with dimension $C \times W' \times H'$, where $C = 512$, $W' = \lfloor \frac{W}{16} \rfloor$, and $H' = \lfloor \frac{H}{16} \rfloor$.

The author use this method to provide potential areas for concealing regions, drawing inspiration from the Region Proposal (Johnson et al., 2016), which employs patches and anchor boxes to construct comprehensive descriptions. The author utilize feature maps within my work 7×7 grid patches to ensure coverage over even small objects in the image. Four parameters are specified for each patch: its width w_p, height h_p, and center position (x_p, y_p). Using a regressive offset model represented by the following equations, the author define k anchor boxes within each patch with four parameters for size (w_a, h_a) and the center position (x_a, y_a) of each anchor box's region:

$$\begin{aligned} x_a &= x_p + t_x \frac{w_p}{2} && (4.1) \\ y_a &= y_p + t_y \frac{h_p}{2} && (4.2) \\ w_a &= w_p \cdot \exp(t_w) && (4.3) \\ h_a &= h_p \cdot \exp(t_h) && (4.4) \end{aligned}$$

In this case, the normalized offset from the center of the anchor is indicated as (t_x, t_y), and the log-scale transformation of the anchor size is represented by $t_w, t_h)$. The author then map (x_a, y_a, w_a, h_a) onto the input feature map's $X \times Y$ grid and then transform the feature map back to its original dimensions of $W \times H$.

4.4.4 Retarget Layer

In addition to the information derived from the feature map for the self-attention process, the author acquire a collection of anchors $R(x_a, y_a, w_a, h_a)$ from the localization layer to use as candidates. $I \in \mathbb{R}^{B \times C \times W \times H}$ is the feature map that the author use, in which B stands for batch size, C for number of channels, W for width, and H for height. The attention map A is first calculated as follows:

$$A(Q, K) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{4.5}$$

In this section $Q, K \in \mathbb{R}^{B \times \frac{C}{8} \times N}$ refers to the query and key matrices that are obtained from the feature map. The scaling factor for the dimensionality of the crucial vectors is determined $n = 8$, is denoted by the term $d_k = \frac{C}{n}$.

Subsequently, the author convert the attention map A to a single dimension of $N = W \times H$ using the value of the feature map matrix V^T. The output feature map is obtained by multiplying the attention map A by V. The output of masking region O is then obtained by scaling this result by a learnable parameter γ and adding the input feature map I, which includes a residual connection as follows:

$$O(A, V, I) = \gamma \cdot Reshape(AV^T) + I \tag{4.6}$$

After that, the output is reshaped using the Reshape() function to return it to the original spatial dimensions, $W \times H$. Then, O is subjected to the sigmoid function to determine the optimal masking anchor boxes M.

$$M = Sigmoid(O) \odot R \tag{4.7}$$

The anchor box candidates R in this case have dimensions of $X \times Y \times 4k$. As a form of optimal masking, the author subsequently remapped M from its $X \times Y \times k$ dimension back to the original images' $X \times Y$ dimension. It is important to notice that, for each anchor parameter, the author apply a mask over the original photos using a masking value M of 1, which is indicated by the condition Sigmoid(O) $>$ 0.6. The author designate areas that are not masked by assigning a value of 0 to the mask, which is indicated by the condition Sigmoid(O) $<=$ 0.6. Next, the author inputted the masked images into the VED framework (Dosovitskiy et al., 2020; Rothe et al., 2020; Li et al., 2021) in order to generate captions C_{gen}.

4.4.5 Loss Function

During the training phase, the author utilize the reference captions C_{true} associated with each image as the ground truth. The author utilize a smooth L1 loss, denoted as $L1_1^{reg}$, in the altered coordinate space (Tanatipuknon et al., 2021) to measure the similarity between the generated captions and the ground truth captions. To assess the similarity of the captions, the author also utilize the inverse cosine similarity loss L_{invc}.

Calculating loss requires an encoding and embedding technique. Using the BERT model, which is known for its effectiveness in creating contextual embeddings, the author build captions C_{gen} based on the reference captions C_{true}. The author utilized the BERT tokenizer to construct encoded vectors for both C_{gen} and C_{true}. Subsequently, the author generate the embeddings V_{gen} and V_{true} by calculating the average of the final hidden states of the encoded vectors in the following manner:

$$\begin{aligned} V_{gen} &= Mean(BERT(Encode(C_{gen}))) \quad (4.8) \\ V_{true} &= Mean(BERT(Encode(C_{true}))) \quad (4.9) \end{aligned}$$

Table 4.2: The statistics of our dataset.

Data source	**Training**	**Validation**	**Testing**	**Total**	**Type**
BGS (bgs, 2023)	1,690	200	146	2,036	Disaster
Normal (lan, 2020)	3,500	669	150	4,319	Normal
DID (did, 2019)	550	70	114	734	Disaster
Shipborne (Li et al., 2023)	-	-	270	270	Disaster and Normal
Total	5,740	939	680	**7,359**	

After obtaining the embedding V_{gen} and V_{true}, the author used them to calculate the inverse cosine similarity loss, L_{invc}, as follow:

$$L_{invc} = \frac{V_{gen} \cdot V_{true}}{||V_{gen}||||V_{true}||} \tag{4.10}$$

The author additionally compute the smooth L1 loss L_1^{reg} in the coordinate space of the transformation, using the following method:

$$L_1^{reg} = \sum_{p \in Parameters} ||p||_1 \tag{4.11}$$

Finally, during the training phase of our model, the author employ a customized loss function that merges the cosine similarity loss L_{invc} and the smooth L1 loss L_1^{reg}. Each of these losses is assigned weights β and α correspondingly. Here is the equation for combined loss:

$$L_{custom} = \alpha \cdot L_{invc} + \beta \cdot L_1^{reg} \tag{4.12}$$

Table 4.3: The statistics of our dataset in each disaster.

Scene type	**Training**	**Validate**	**Testing**	**Source**
Landslide	1,690	200	416	BGS, DID, Shipborne
Flood	225	35	60	DID
Fire	225	35	54	DID
Normal	3,500	669	150	Kaggle

4.5 Experiments and results

Our approach involves using the RetNet algorithm to analyze photos and generate region proposals, which are then inputted into RetNet. This section presents the results of our proposed method, including its performance with different ablation loss functions, patch grid counts, and anchor box counts. Furthermore, the author utilize our network to examine additional calamities such as floods and wildfires, and produce descriptive titles for the different situations. In addition, the author employ our system to examine landslide situations using shipborne imagery, and evaluate its effectiveness in comparison to traditional detection methods.

4.5.1 Dataset

In this study, the author utilize image datasets from four primary sources, as detailed in Table 4.2:

- The British Geological Survey (BGS) (bgs, 2023) provided the landslide images. The author further annotated these images with text captions to enhance our dataset for the intended analyses.
- Kaggle (lan, 2020) contributed an extensive collection of 4,319 images, from which derived the common scene image dataset.
- Disaster Image Dataset (DID) (did, 2019) provided the flood and wildfire images. The author annotated the images with captions for our network.
- Shipborne (Li et al., 2023) provided the landslide and non-landslide images that were captured during a survey conducted from a ship.

The models in our system were trained, validated, and tested utilizing data sets from BGS, Kaggle, and DID. The Shipborne dataset was exclusively utilized for the purpose of testing in a classification application. Table 4.2 indicates that there are a total of 2,036 abnormal images from BGS and 4,319 typical images from Kaggle. In addition, the author expanded the disaster datasets to include floods and wildfires, using the methodology described in the DID report (did, 2019). The author have chosen to incorporate these two parts on catastrophic events within our studies in order to examine the different scenarios. Table 4.3 demonstrates the dataset the author used in each disaster type; there are consists of Landslide, Flood, Wildfire for disaster type, and Normal for non-disaster type from Kaggle.

The author made a deliberate attempt to incorporate often used label terms from this collection, such as trees, rocks, soil, water, rivers, fires, and lakes. A textual caption was added to the dataset. In addition, the author utilized the terminology of the objects themselves to depict the surrounding objects and their respective positions. This approach was employed to enhance the model's comprehension of the scene's circumstances.

The author utilized the shipborne image-based landslide dataset (Li et al., 2023), which has a total of 270 photographs. Among these, 231 images are unrelated to landslides, while the remaining 39 images depict landslides. Consequently, the author was able to apply the categorization to a different dataset. This dataset is utilized as a benchmark to assess the efficacy of our suggested methodology.

4.5.2 Ablation Study

Our methodology utilizes image masking to redirect attention throughout the process of generating captions for images. The L1 regularization loss function proposed by Girshick et al.(Girshick, 2015) is recommended for optimizing parameters, specifically for anchor box form. To enhance the retarget layer and ensure that the captions are equivalent to the labels, the author utilize a cosine similarity loss function. The author developed an inverse cosine similarity loss method to identify distinct captions. This approach promotes the masking of images to emphasize specific objects, which is innovative in the identification of distinct objects using an inverse function. Our collection includes images and captions that showcase a diverse range of things unrelated to disasters, such as "The man in front of the rocky mountain, whose soil has collapsed."

Table 4.4: The ablation of loss functions in our proposed method.

Cosine	Inv Cosine	L1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
✓	–	–	0.360	0.276	0.234	0.199	0.224	0.350	1.260
–	✓	–	0.378	0.295	0.253	0.218	0.237	0.362	1.470
–	–	✓	0.415	**0.339**	0.297	**0.258**	**0.270**	0.400	1.770
✓	–	✓	0.415	**0.339**	0.297	**0.258**	**0.270**	0.400	1.770
–	✓	✓	**0.416**	**0.339**	**0.298**	**0.258**	**0.270**	**0.401**	**1.779**
✓	✓	✓	**0.416**	**0.339**	**0.298**	**0.258**	**0.270**	**0.401**	**1.779**

The L1 regularization loss function has a considerable impact on the model's performance, as demonstrated by the results presented in Table 4.4. BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) are standard linguistic metrics the author utilize to assess the relevance of text captions.The findings indicated that the inverse cosine similarity loss function outperforms the cosine similarity loss function. The combination of inverse cosine similarity and L1 regularization produced the following scores for several evaluation metrics: 0.416 for BLEU-1, 0.339 for BLEU-2, 0.298 for BLEU-3, 0.258 for BLEU-4, 0.270 for METEOR, 0.401 for ROUGE-L, and 1.799 for CIDEr. While the inverse cosine similarity performs marginally better than the cosine similarity loss, incorporating an additional loss function does not improve the results beyond the combination of inverse cosine similarity and L1 regularization.

Table 4.5: The ablation of number of patch grid with 3 anchor boxes in our proposed method.

Number of patch grid	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
3	0.416	0.339	0.298	0.258	0.270	0.401	1.779
4	0.416	0.339	0.298	0.258	0.270	0.401	1.779
5	0.416	0.339	0.298	0.258	0.270	0.401	1.779
7	0.416	0.339	0.298	0.258	0.270	0.401	1.779

Table 4.6: The ablation of number of anchor boxes with 7 patch grids in our proposed method.

Number of anchor boxes	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
3	0.416	0.339	0.298	0.258	0.270	0.401	1.779
5	0.416	0.339	0.298	0.258	0.270	0.401	1.779
7	0.416	0.339	0.298	0.258	0.270	0.401	1.779

In addition, as shown in Table 4.5, the author conducted experiments with patch grids of sizes 3, 4, 5, and 7, with each grid containing three anchors.

In addition, the author conducted experiments with seven patch grids, using three, five, and seven anchor boxes, in order to determine the configurations that optimize the performance of the model. The results are presented in Table 4.6. Although the number of anchor boxes in each grid and the patch grids were adjusted, these alterations had no meaningful effect on the model's performance as measured by all evaluated metrics. The scores for BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr remained constant at 0.416, 0.339, 0.298, 0.258, 0.270, 0.401, and 1.799, respectively.

4.5.3 Flood and Wildfire

The author refined the model by utilizing photographs of floods and wildfires collected from the Disaster Image Dataset (DID). Descriptive captions were added to each image to be used during both the training and testing phases. The findings suggest that the model exhibits superior performance in identifying and describing photographs of wildfires compared to images of floods, as depicted in Figure 4.7. The BLEU-1 score for photographs depicting floods is precisely 0.223, and for images portraying fires, it is precisely 0.205. The ROUGE-L score for floods, which is 0.230, is somewhat higher than the score for wildfires, which is 0.219. This indicates a better similarity in the longest common subsequences found in the captions of flood photographs. In addition, the METEOR scores of 0.146 for floods and 0.156 for wildfires demonstrate a satisfactory level of semantic and grammatical agreement with the captions supplied by the reference. The marginally elevated score for wildfires indicates improved model accuracy in depicting wildfire imagery. Furthermore, the METEOR and CIDEr ratings of 0.423 for floods and 0.627 for wildfires indicate that the captions for wildfire images are more accurate and closely match the reference evaluations.

Table 4.7: Performance metrics for landslide, flood, and wildfire disaster in image captioning.

Disaster	**BLEU-1**	**METEOR**	**ROUGE_L**	**CIDEr**
Landslide	0.416	0.270	0.401	1.779
Flood	0.223	0.146	0.230	0.423
Wildfire	0.205	0.156	0.219	0.627

4.5.4 Classification

The classification performance of RetNet, VED, and ResNet50 (Ofli et al., 2021) has been compared and the results are presented in Table 4.8. In this classification experiment, the author used two datasets from different viewpoints: side view from the BGS dataset and shipborne view from the Shipborne dataset. For the BGS dataset, the author used the testing set, which was split from the dataset. However, the Shipborne dataset was not used for training, but only for the inference process to evaluate the performance of RetNet with unseen data. Based on the recall, accuracy, precision, and F1-score values of 0.9160, 0.9067, 0.9510, and 0.9283, respectively, it is evident that our strategy outperforms the others. In addition, the author utilized data obtained from surveys conducted on ships to assess the accuracy of the classification process from a new perspective. RetNet's accuracy, although lower than that of the Fusion approach (Li et al., 2023), surpassed the Fusion approach in terms of F1-score, recall, and precision. Significantly, our method surpassed Fusion in performance, with a recall rate of 0.9643 compared to Fusion's 0.8290.

Table 4.8: Classification performance on different datasets.

Method	**BGS Dataset**		**Shipborne Dataset**	
	RetNet	**Ofli et al.**	**RetNet**	**Li et al.**
Accuracy	**0.9160**	0.8700	0.8750	**0.9444**
Precision	**0.9067**	0.7370	**0.8852**	–
Recall	**0.9510**	0.6680	**0.9643**	0.8290
F1-Score	**0.9283**	0.7010	**0.9231**	–

4.6 Discussion

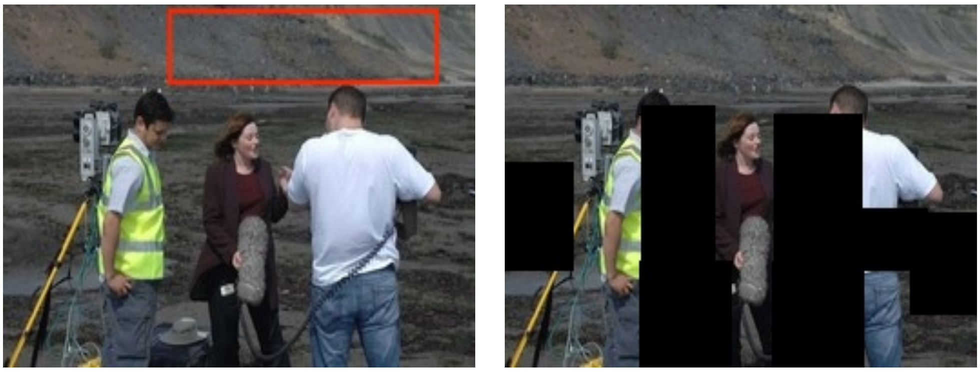
4.6.1 Image Captioning

The findings indicate that both the size of the patch and the quantity of anchor boxes did not have a significant impact on the produced captions, as demonstrated in Table 4.4. The localization layer's region candidates were considered as potential options for selection as the ideal ones for masking.

The patch size and number of anchor boxes are not significant factors in image captioning because the masked sections often occupy comparable places.

In relation to the side-view images depicted in Figure 4.6 and Figure 4.7, our model effectively prioritizes the target objects, unlike the original captions which may have focused on or misread other features. Moreover, the Figure 4.8 and Figure 4.9 demonstrate the encouraging outcome of landslide detection using shipborne images. Similarly, the aerial images depicted in Figure 4.10 and Figure 4.11 showcase the model's proficiency in detecting landslides from an elevated viewpoint. Despite the model being mostly trained on side-view photographs, our approach surpasses other methods in detecting landslides in aerial view images.

Figure 4.6 shows the image containing the persons in front of the image, while behind the scene, there is a disaster region and a landslide at the top in the background. Figure 4.6 (right) shows the result of RetNet, which masked some part of the human to shift the attention of the image captioning Vision Encoder-Decoder model to detect the disaster region. The original image caption demonstrates that the people investigating the damaged ground which is not specific to the landslide region at the back. After masking the image from RetNet, the masked caption represents more specific detail in the scene without completely masking the major objects, humans, and instruments.

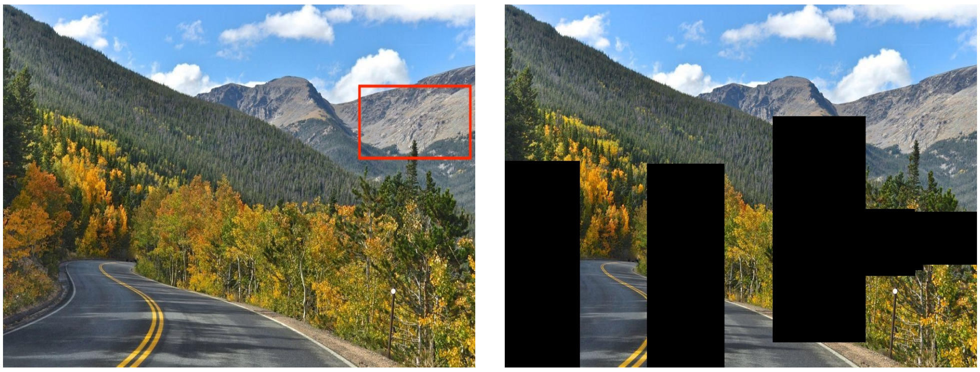


people investigate the damaged ground

Collapsed hill caused by soil slide

Figure 4.6: The original image caption and masked image caption to detect disaster at the background.

Figure 4.7 demonstrates the original image caption (left) that could detect and explain the soil and rocks sliding down to the road. However, the soil and rocks slide down from the hill/mountains but do not fall to the road. Applying RetNet to mask some parts of the image before generating a caption is much better than an original caption in this situation. Nevertheless, the masked caption in Figure 4.7 (right) contains *logs*, which seem not to appear in the scene.

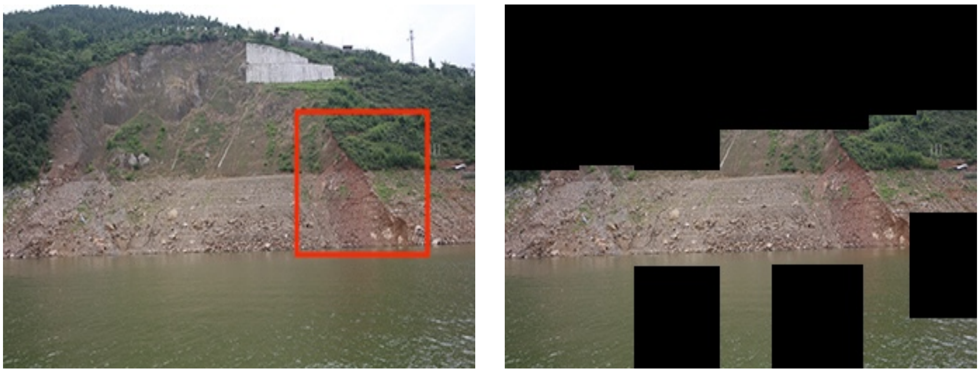


soil and rocks slide down to the road

logs slide down from the hill

Figure 4.7: The original image caption and masked image caption without major object in the scene.

The original and masked image caption could detect the disaster in the scene in Figure 4.8. The caption from the original image Figure 4.8 (left) seems to explain the detail better than the caption from the masked image. While the masked caption in Figure 4.8 (right) attention to the correct region of the image with no specific caption.

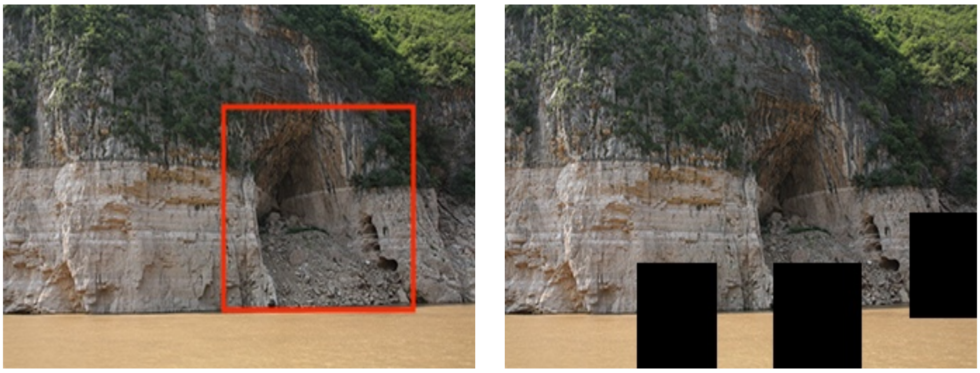


damaged hill by soil slide

cracked rocks and ground

Figure 4.8: Side view from shipborne detecting the landslide.

Figure 4.9 demonstrates the complex scene for a computer vision model in which the landslide blends to the rocky cliff from a shipborne side view image. The caption from an original image cannot detect the disaster with the rocky cliff; detect the rocky cliff and water body in Figure 4.9 (left). The RetNet masked caption can completely detect the landslide disaster at the rocky cliff in Figure 4.9 (right) by masking some part of the image between the rocky cliff and the water body.

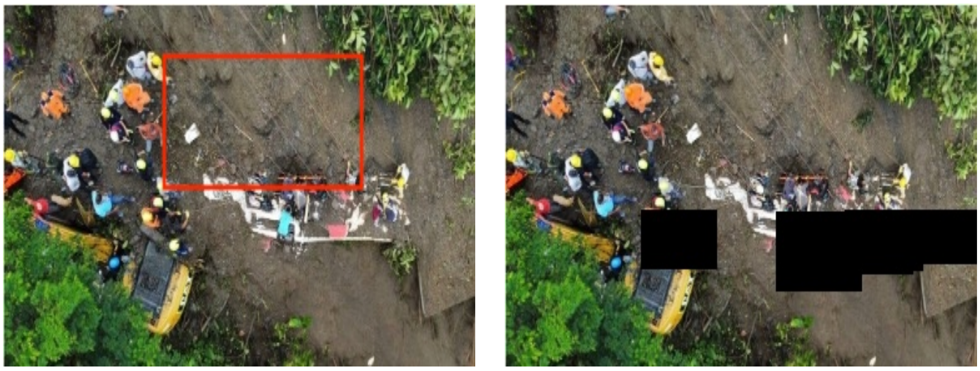


rocky cliff in front of sea

damaged cliff by soil slide

Figure 4.9: Shipborne image with landslide blend to the cliff.

Figure 4.10 (left) shows the caption from the original image explaining the obstacles to the snow in the mountain. The RetNet caption in Figure 4.10 (right) demonstrates the correct caption with soil and rocks sliding down from the hill. The masking region at the obstacle (the white particles in the scene) could achieve the caption, making the classification part more accurate.

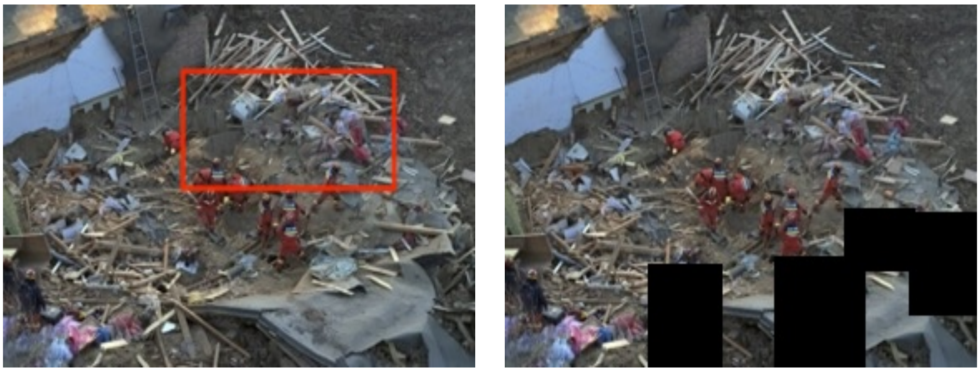


snow capped mountain range

soil and rocks slide down from hill

Figure 4.10: Aerial image view at the landslide region with the obstacle things.

The caption from the original image and the masked image generated the correct situation in the scene and were quite similar to each other in Figure 4.11. In this image, the disaster scene contains landslides and other objects without the typical region. The model's attention to any part of the image could detect the disaster region, which made the caption from the original image and RetNet not different.

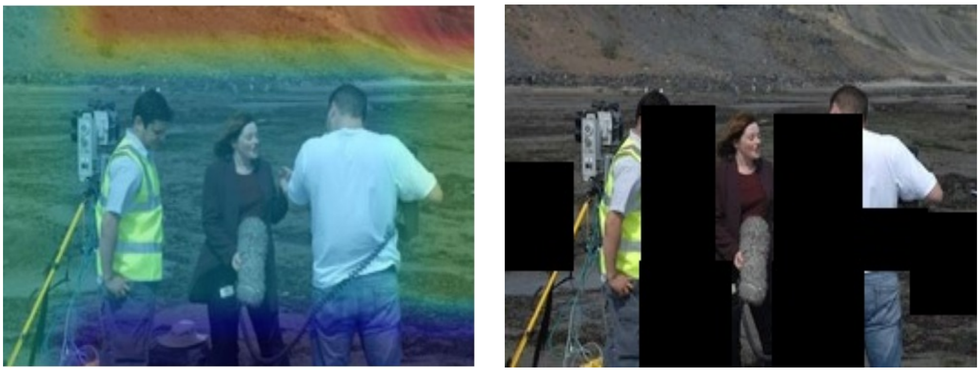


collapsed hill caused by fallen soil

destroyed hill caused by fallen soil

Figure 4.11: The aerial view with complex objects in the scene.

Additionally, the author present a heat map that illustrates the new attention derived from RetNet. For landslide disaster images, as seen in Figure 4.12, the heat map indicates that our model shifts its focus from people to the landslide region in the upper part of the image, thereby generating captions that describe the disaster situation in the targeted region.

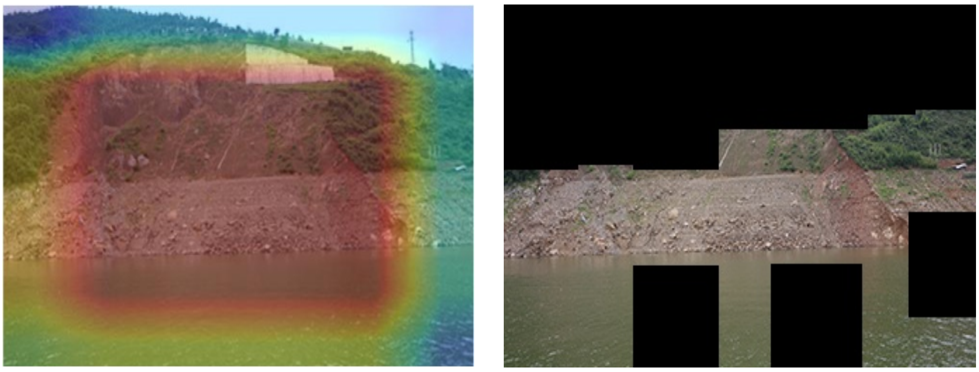


People investigate the damaged ground

Collapsed hill caused by soil slide

Figure 4.12: The landslide from side view with attention color map.

Conversely, in Figure 4.13, where the landslide occupies the center of the image without being obscured by other objects, the application of masking via RetNet enables the VED to concentrate exclusively on the specific landslide region depicted in the image.

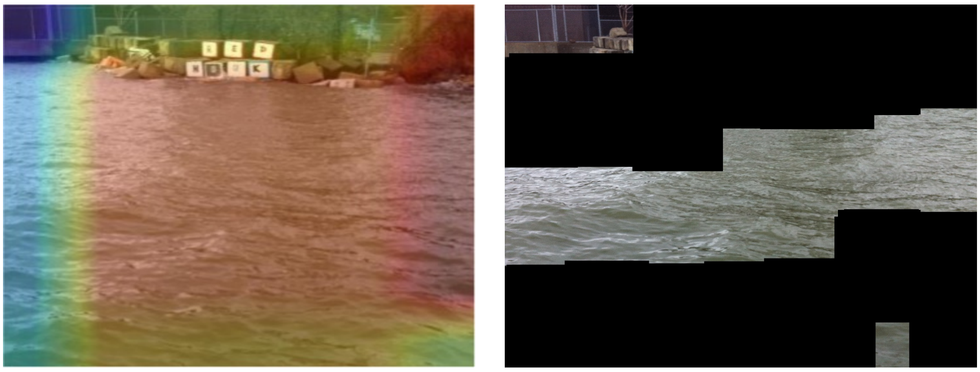


Damaged hill by soil slide

Cracked rocks and ground

Figure 4.13: The landslide from shipborne view with attention color map.

In scenarios involving floods, the captions generated by our model accurately identify the specific areas depicted in heat map images (Figure 4.14).

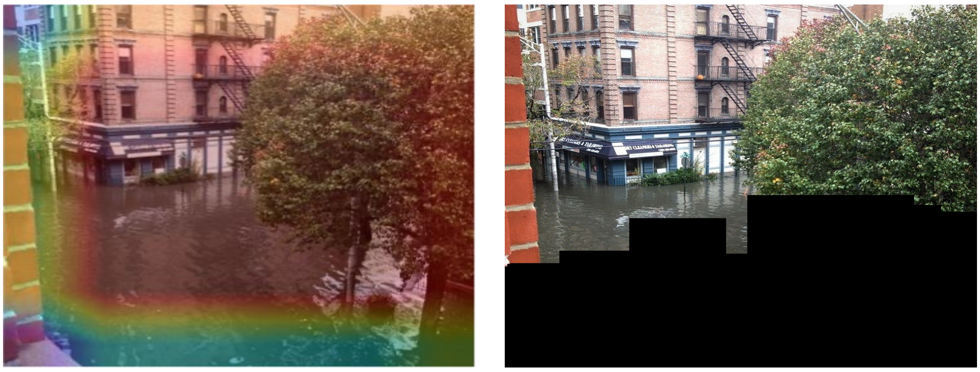


Car driving through water on the road

Rapids flow in a sea

Figure 4.14: The flood from side view without other objects in the scene and attention color map.

However, it is common for the original image (as shown in Figure 4.15 (left)) and the corresponding masked image (as illustrated in Figure 4.15 (right)) to produce identical captions. This occurrence is due to the extensive water regions in the images, which are sizable enough to be detected without the need for attention to shift.

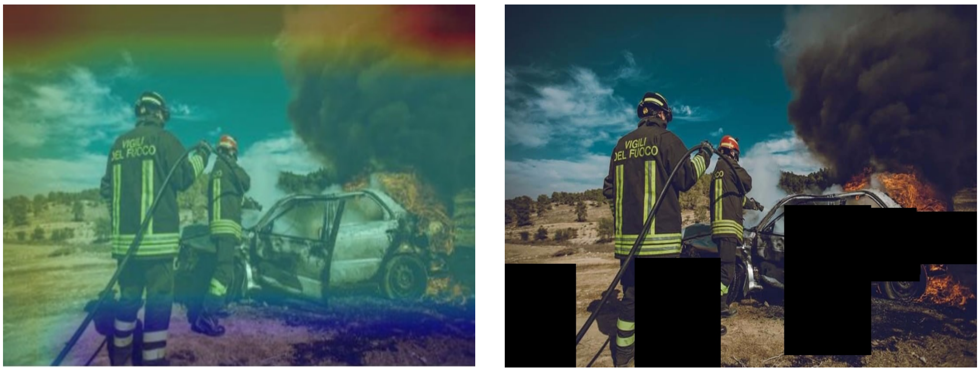


Water overflow into a town

Water overflow into a town

Figure 4.15: The flood from side view which flood is in the center of image and attention color map.

For the wildfire disaster, Figure 4.16 (left) and 4.16 (right) show that our model not only allow VED to focus on wildfire but also other surrounding objects and humans. As a result, generated captions are correct even it is from different aspects from the original image.

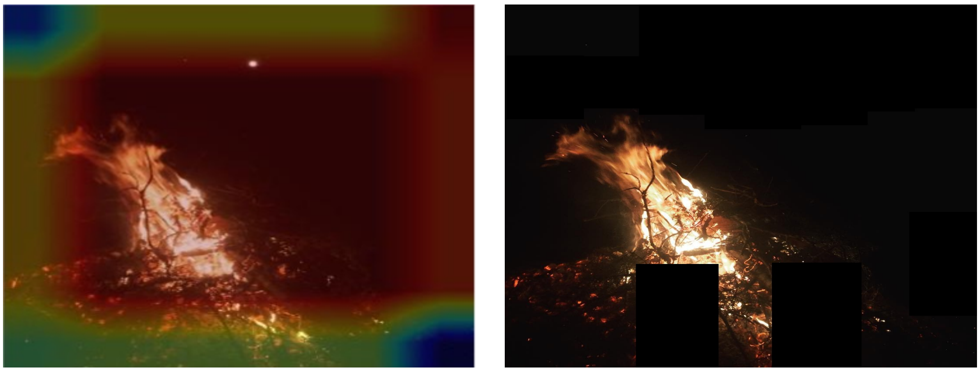


Smoke covers the fire area

Fireman controls the fire

Figure 4.16: Fire disaster with complex objects and attention color map.

Moreover, in Figure 4.17, the original image caption misinterpret the phenomena. On the other hand, our model could generate the correct perspective, but only partially correct captions as shown in Figure 4.17. In general, the author realized that the VED captioning model would pay attention to the center of the image to generate the circumstance captions, while RetNet captions attention to the specific regions.

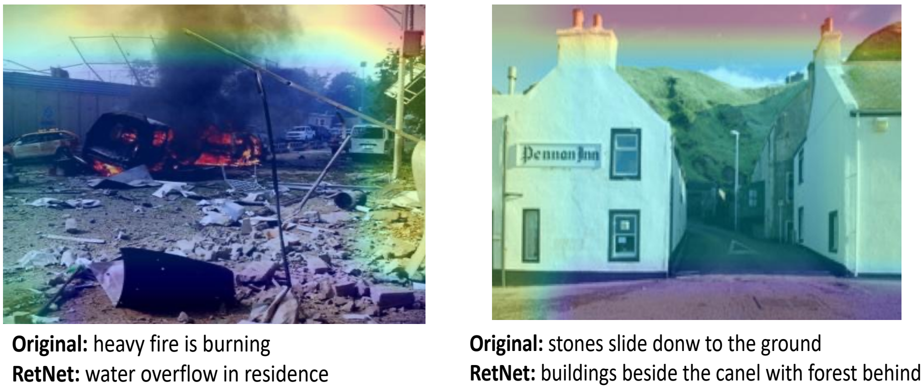


Rapids flow in canal

Fireman spraying water to put out fire

Figure 4.17: The wildfire from side view in dark scene with attention color map.

Finally, the author showcase instances of mistaken recognition by proposed method in Figure 4.18. RetNet erroneously directs attention towards the upper sections of the lateral perspective of a fire catastrophe image (as depicted in Figure 4.18 (left)), instead of the central area. The misguided emphasis leads to the production of captions that are not accurate. Similarly, in the lateral perspective image of a landslide (as depicted in Figure. 4.18 (right)), RetNet has a tendency to concentrate on the upper portion of the image instead than the accurate target areas situated in the center.

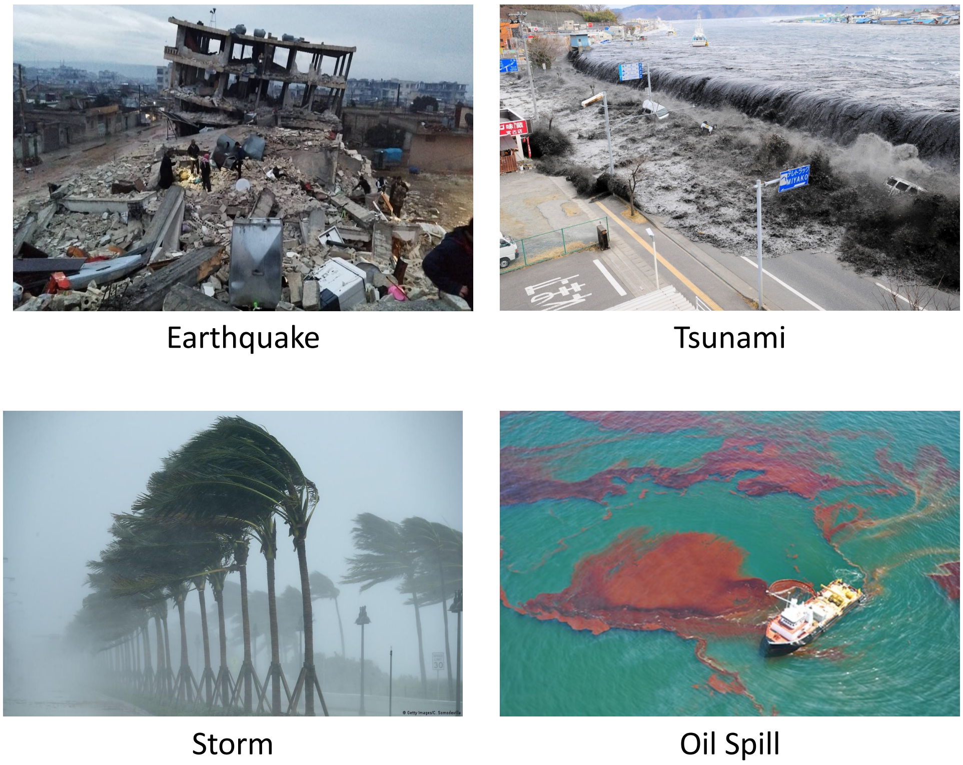

Original: heavy fire is burning
RetNet: water overflow in residence
Pennan Inn
Original: stones slide donw to the ground
RetNet: buildings beside the canel with forest behind

Fire disaster failure case.

Landslide disaster failure case.

Figure 4.18: Heat map attention of caption from original image and RetNet image in failure case.

RetNet could be used for other types of disasters, as represented in Figure 4.19. Earthquake events destroy constructions such as buildings, roads, or highways, leaving debris over urban areas. In this case, RetNet could be used to detect debris scenes as well. Similarly, in a tsunami event, the aftermath



Earthquake

Tsunami

Storm

Oil Spill

Figure 4.19: Other type of disaster such as earthquake, tsunami, storm, and oil spill.

involves flooding and debris flows, which are similar to flood and landslide disasters. However, storm disasters present some limitations, as they are quite difficult to detect. Since wind is not visible in RGB images and the effect of storms may only cause slight movements in objects like trees, detection is challenging. In a similar vein, oil spills can be detected by RetNet through color changes in the ocean. However, as shown in Figure 4.19, oil spills sometimes appear similar to land.

4.6.2 Classification

Our technique focuses on specific regions, as indicated by Table 4.8, which demonstrates that RetNet achieves better results than Ofli et al. (Ofli et al., 2021) in analyzing side view images, particularly with the BGS dataset. Contrasting with the findings of Li et al. (Li et al., 2023), our model demonstrates reduced rates of accurate positive and accurate negative predictions when utilized with shipborne photos. However, RetNet outperforms other methods in terms of recall, indicating its enhanced ability to identify landslides in unclear situations.

4.7 Conclusion

Detecting disaster-related locations in aerial or shipborne imagery is typically challenging due to their small size within the photos. This research presents a new framework called the Retarget Network (RetNet), which aims to improve the capability of image-captioning-based machine learning models to identify important regions in an image. The network we propose has a distinct feature of adjusting detection priorities by combining a localization layer with a retarget layer, using a combination of patch and anchor box techniques. The RetNet model underwent comprehensive testing over a spectrum of disaster situations, encompassing landslides, floods, and wildfires, from diverse viewpoints. The results of our study indicate that by preprocessing images using RetNet and then analyzing them with a Vision Encoder-Decoder (VED), the accuracy of landslide detection in side-view image captions improves significantly to 91.60%. Additionally, for images captured from shipborne perspectives, the accuracy rate achieved is 87.50%.

4.8 Contribution

The contributions of the work presented in this chapter were:

- Inspired by the state-of-the-art anchor boxes, this approach applies masked regions. In this part, masking is used to exclude regions that are not significant for detection, thereby retargeting the model's attention to the areas of interest.
- Apply the optimal equations to generate the size and position of anchor boxes.
- Design the model to achieve two main objectives: generating masked candidate anchor boxes and selecting the optimal masked regions using a loss function derived from captions.

Chapter 5

Conclusion

This chapter will concludes the study by addressing the significant research findings related to the research questions, philosophies, and contributions. It will also discuss the study's limitations and provide suggestions for further research.

5.1 Summary of Major Findings

This study aims to detect disaster scenes within environmental scenes and natural objects. Conventional techniques often fail to achieve this objective accurately due to the limitations of pixel-based methods. As a result, this study successfully distinguishes between natural and disaster objects. Furthermore, in complex scenes where other objects, such as humans, dominate the center of the image and disaster regions are in the background or too small compared to other objects, the challenge is significant. The results demonstrated that the proposed architecture in this study can shift and retarget the model's attention to significant objects, namely the disaster regions.

In Chapter 3, the proposed method classifies disaster scenes from natural scenes using text-based techniques instead of relying solely on pixel-based methods. The methodology employs Vision Encoder Decoder (VED) techniques to encode the image, capturing the relationship between it and its caption. The results showed that the proposed method achieved an accuracy, precision, recall, and F1-score of 95.00%, 96.19%, 96.42%, and 96.31%, respectively. In comparison, the conventional method, ResNet50, achieved an accuracy of 71.61%, precision of 98.81%, recall of 70.77%, and F1-score of 82.47%. In AUC analysis, the proposed method achieved 94% accuracy in distinguishing between normal and landslide images. In this study, the ResNet50 model trained from scratch proved more effective at classifying landslide images than the ResNet50 model fine-tuned using ImageNet. Therefore, the results demonstrate that the caption-based method provides more information for classifying scenes than pixel-based methods alone. Furthermore, the information from human-language captions captures significant features and action details, offering more comprehensive insights into the actions and relationships within the scene through text tokens instead of pixel tokens.

In Chapter 4, the proposed architecture, Retarget Network (RetNet), uses state-of-the-art techniques inspired by anchor boxes to generate masked candidate regions. Simultaneously, the features extracted from the image are used to select the optimal masked positions, enabling accurate disaster detection. In this study, the proposed method employs VGG16 for feature extraction before generating anchor boxes, which serve as candidate masking regions. In this process, the shape and center of each anchor position are set according to specific conditions within each patch. After generating the anchor boxes, the optimal ones are selected to effectively mask the areas, shifting the model's attention to detect the target regions. From diverse viewpoints, the RetNet model underwent comprehensive testing across various disaster situations, including landslides, floods, and wildfires. The results indicate that preprocessing images using RetNet, followed by analysis with a Vision Encoder-Decoder (VED), significantly improves the accuracy of landslide detection in side-view image captions to 91.60%. Additionally, the accuracy rate achieved for images captured from shipborne perspectives is 87.50%.

Referring to the research questions and philosophy, the author aims to study and develop a machine-learning model capable of detecting disaster scenes from various perspectives without relying on boundary labeling. The image captioning features discussed in Chapter 3 provide detailed and explanatory insights to address the research questions. Moreover, conventional methods tend to focus on the major objects in complex scenes where disaster objects coexist with other objects. However, the proposed method can shift attention, prioritize, and optimize the model to achieve the desired objectives. The results of this study demonstrate that the proposed method can adapt to different viewpoints in images and can be applied to various real-world scenarios. This adaptability highlights the potential for broader application beyond this research's specific disaster detection scenarios.

However, during the experiment, the limitation of caption tokens led to insufficient information for detection in some cases. Additionally, the proposed method sometimes needed help to generate accurate captions when the situation in the scene was on a small scale and could not be classified as a disaster. There is also a limitation in retargeting the detection when there are more than two major objects in the scene, which can cause the RetNet to fail in detecting the disaster area.

5.2 Suggestions for Future Work

The architecture could be improved by expanding the number of generated text tokens from 30. Enhancing the caption generation component with a more robust Natural Language Processing model would provide better information for detection. To improve the performance of the Vision Encoder-Decoder (VED), various techniques can be applied. For the vision part, using methods such as Shift Windows (SWIN) can help to focus more attention. For the caption part, experimenting with different types of tokenizers can generate more concrete captions.

To address the limitation in detection when there are more than two major objects in the scene, implementing a looping inference mechanism to shift attention until the target region is identified could be beneficial. Additionally, incorporating multi-input and multi-modal approaches could provide more comprehensive information, leading to more accurate detection.

Bibliography

(2019). Disaster image dataset.

(2020). Landscape pictures.

(2023). The national archive of geological photographs.

Akamine, S., Totoki, S., Itami, T., and Yoneyama, J. (2022). Real-time obstacle detection in a darkroom using a monocular camera and a line laser. *Artificial Life and Robotics*, pages 1–6.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Can, R., Kocaman, S., and Gokceoglu, C. (2019). A convolutional neural network architecture for auto-detection of landslide photographs to assess citizen science and volunteered geographic information data quality. *ISPRS International Journal of Geo-Information*, 8(7):300.

Castro, R., Pineda, I., Lim, W., and Morocho-Cayamcela, M. E. (2022). Deep learning approaches based on transformer architectures for image captioning tasks. *IEEE Access*, 10:33679–33694.

Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R., Fua, P., Salzmann, M., and Rottmann, M. (2021). Segmentmeifyoucan: A benchmark for anomaly segmentation. *arXiv preprint arXiv:2104.14812*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhruv, P. and Naskar, S. (2020). Image classification using convolutional neural network (cnn) and recurrent neural network (rnn): a review. *Machine Learning and Information Processing: Proceedings of ICMLIP 2019*, pages 367–381.

Di Biase, G., Blum, H., Siegwart, R., and Cadena, C. (2021). Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Haseeb, M., Xinhailu, A. B., Khan, J. Z., Ahmad, I., and Malik, R. (2011). Construction of earthquake resistant buildings and infrastructure implementing seismic design and building code in northern pakistan 2005 earthquake affected area. *International Journal of Business and Social Science*, 2(4).

Hernández, D., Cecilia, J. M., Cano, J.-C., and Calafate, C. T. (2022). Flood detection using real-time image segmentation from unmanned aerial vehicles on edge-computing platform. *Remote Sensing*, 14(1):223.

Hungr, O., Leroueil, S., and Picarelli, L. (2014). The varnes classification of landslide types, an update. *Landslides*, 11:167–194.

Ibrahim, N., Sharun, S., Osman, M., Mohamed, S., and Abdullah, S. (2021). The application of uav images in flood detection using image segmentation techniques. *Indones. J. Electr. Eng. Comput. Sci*, 23(2):1219.

Johnson, J., Karpathy, A., and Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

Kamoi, R., Iida, T., and Tomite, K. (2021). âefficient unknown object detection with discrepancy networks for semantic segmentation. In *Proc. NeurIPS Workshop Mach. Learn. Auton. Driving*, pages 1–12.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). 2012 alexnet. *Adv. Neural Inf. Process. Syst*, pages 1–9.

Li, H., He, Y., Xu, Q., Deng, J., Li, W., and Wei, Y. (2022). Detection and segmentation of loess landslides via satellite images: A two-phase framework. *Landslides*, 19(3):673–686.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2021). Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Li, Y., Wang, P., Feng, Q., Ji, X., Jin, D., and Gong, J. (2023). Landslide detection based on shipborne images and deep learning models: a case study in the three gorges reservoir area in china. *Landslides*, 20(3):547–558.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lis, K., Nakka, K., Fua, P., and Salzmann, M. (2019). Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161.

Liu, P., Wei, Y., Wang, Q., Chen, Y., and Xie, J. (2020). Research on post-earthquake landslide extraction algorithm based on improved u-net model. *Remote Sensing*, 12(5):894.

Meena, S. R., Soares, L. P., Grohmann, C. H., Van Westen, C., Bhuyan, K., Singh, R. P., Floris, M., and Catani, F. (2022). Landslide detection in the himalayas using machine learning algorithms and u-net. *Landslides*, 19(5):1209–1229.

Mirzaev, I., Yuvmitov, A., Turdiev, M., and Shomurodov, J. (2021). Influence of the vertical earthquake component on the shear vibration of buildings on sliding foundations. In *E3S Web of Conferences*, volume 264, page 02022. EDP Sciences.

Nefeslioglu, H. A., Duman, T. Y., and Durmaz, S. (2008). Landslide susceptibility mapping for a part of tectonic kelkit valley (eastern black sea region of turkey). *Geomorphology*, 94(3-4):401–418.

Ofli, F., Imran, M., Qazi, U., Roch, J., Pennington, C., Banks, V. J., and Bossu, R. (2021). Landslide detection in real-time social media image streams. *arXiv preprint arXiv:2110.04080*.

Ofli, F., Qazi, U., Imran, M., Roch, J., Pennington, C., Banks, V., and Bossu, R. (2022). A real-time system for detecting landslide reports on social media using artificial intelligence. In *Web Engineering: 22nd International Conference, ICWE 2022, Bari, Italy, July 5–8, 2022, Proceedings*, pages 49–65. Springer.

Ohgushi, T., Horiguchi, K., and Yamanaka, M. (2020). Road obstacle detection method based on an autoencoder with semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*.

Pan, H., Badawi, D., and Cetin, A. E. (2020). Computationally efficient wildfire detection method using a deep convolutional network pruned via fourier analysis. *Sensors*, 20(10):2891.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pennington, C. V., Bossu, R., Ofli, F., Imran, M., Qazi, U., Roch, J., and Banks, V. J. (2022). A near-real-time global landslide incident reporting tool demonstrator using social media and artificial intelligence. *International Journal of Disaster Risk Reduction*, 77:103089.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rothe, S., Narayan, S., and Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Sameen, M. I. and Pradhan, B. (2019). Landslide detection using residual networks and the fusion of spectral and topographic information. *IEEE Access*, 7:114363–114373.

Sardogan, M., Tuncer, A., and Ozen, Y. (2018). Plant leaf disease detection and classification based on cnn with lvq algorithm. In *2018 3rd international conference on computer science and engineering (UBMK)*, pages 382–385. IEEE.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical*

Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings, pages 146–157. Springer.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Soares, L. P., Dias, H. C., and Grohmann, C. H. (2020). Landslide segmentation with u-net: Evaluating different sampling methods and patch sizes. *arXiv preprint arXiv:2007.06672*.

Tanatipuknon, A., Aimmanee, P., Watanabe, Y., Murata, K. T., Wakai, A., Sato, G., Hung, H. V., Tungpimolrut, K., Keerativittayanun, S., and Karnjana, J. (2021). Study on combining two faster r-cnn models for landslide detection with a classification decision tree to improve the detection performance. *Journal of Disaster Research*, 16(4):588–595.

Tantanee, S., Wandee, K., and Tovichakchaikul, S. (2018). One page project management application on flood preparedness: Case study of thailand. *Procedia engineering*, 212:363–370.

Thanyawet, N., Ratsamee, P., Uranishi, Y., and Takemura, H. (2023). Abnormal scene classification using image captioning technique: A landslide case study. In *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7. IEEE.

Thomas, R. S. and Hall, B. (2015). *Seawall design*. Butterworth-Heinemann.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.

Trnovszky, T., Kamencay, P., Orjesek, R., Benco, M., and Sykora, P. (2017). Animal recognition system based on convolutional neural network. *Advances in Electrical and Electronic Engineering*, 15(3):517–525.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. (2020). Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.

Zhu, X., Li, L., Liu, J., Peng, H., and Niu, X. (2018). Captioning transformer with stacked attention modules. *Applied Sciences*, 8(5):739.