Detective Networks: Enhancing Disaster Recognition in Images Through Attention Shifting using Optimal Masking (Supplementary Material)

Narongthat Thanyawet[1], Photchara Ratsamee[2], Yuki Uranishi[1] and Haruo Takemura[1]
[1]Osaka University, Osaka, Japan narongthat.thanyawet@lab.ime.cmc.osaka-u.ac.jp
[2]Osaka Institute of Technology, Osaka, Japan. ratsamee.photchara@oit.ac.jp

1. Overview

In this supplementary material, we provide more detailed information about the Masking Candidate layer, to which we have introduced a novel variation from the Region Proposal Network (RPN) [5, 6, 9], as discussed in Section 2 of the manuscript. We provide more detailed information about the datasets used, which include the landslide dataset from the British Geological Survey (BGS) [3] and the non-landslide dataset from Kaggle [2], as outlined in Section 3. Additionally, we conducted ablation studies to investigate the effects of varying anchor boxes, grid patches, and feature selection models, as described in Section 4. Finally, in Section 5, we present and discuss the limitations of this study.

2. Masking Candidate Layer

In this section, we provided more detail explanation of the Masking Candidate layer, which generate the the Candidate Proposals CP as described in Section 4.3 of the manuscript. We inspired from Region Proposal Network (RPN) [5, 6, 9] which generate the anchor box to find the highest Intersection of Union (IoU) between generated boundary and labelled boundary. In Detective Network (DeNet), we used the state-of-the-art of anchor box generating using the feature map I to produce (t_x, t_y) using the hyperbolic tangent (tanh) to ensure that x_a, y_a are in the patch with $\frac{w_p}{2} and \frac{h_p}{2}$ terms, and $t_w, t_h)$ with the rectified linear unit (ReLU) for the anchor boxes size in Figure 1.

$$x_a = x_p + t_x \frac{w_p}{2} \quad (1)$$

$$y_a = y_p + t_y \frac{h_p}{2} \quad (2)$$

$$w_a = w_p \exp(t_w) \quad (3)$$

$$h_a = h_p \exp(t_h) \quad (4)$$

As preliminary experiments, we found that the masked regions shift attention to the fact that there is no need to cover the whole major object. Figure 2 demonstrates an image is patched into 3×3, and the width w_p and the height h_p represent the size of each patch. Each patch has the same size for the center position (x_p, y_p). In Equation 1 and 2, we would like to generate anchor boxes center of position in the patch region. We then used the hyperbolic tangent (tanh) to produce (t_x, t_y). The hyperbolic tangent (tanh) range between $(-1, 1)$ which make Equation 1 and 2 that used $\frac{w_p}{2}, \frac{h_p}{2}$ and multiply by (t_x, t_y) would not exceed the patch boundary.

Moreover, the Equation 3 and 4 use (w_p, h_p) and multiply by exponential of (t_w, t_h) which used the rectified linear unit (ReLU) for the anchor boxes size. The range of the rectified linear unit (ReLU) to produce (t_w, t_h) is $[0, \infty)$, which makes the size of anchor boxes cover the objects.

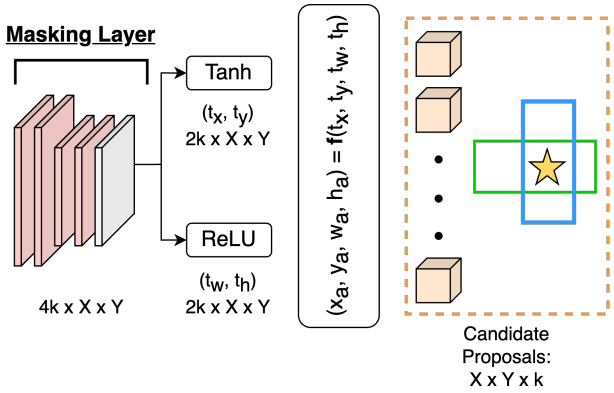

Figure 1. Masking Candidate Layer.

3. Dataset contribution

The dataset from the British Geological Survey (BGS) [3] and the Kaggle [2] that we labeled into the 30 tokens caption in each image. Table 1 demonstrates the dataset we used in the training and inference process. In landslide images, we use the images from the BGS dataset in the training

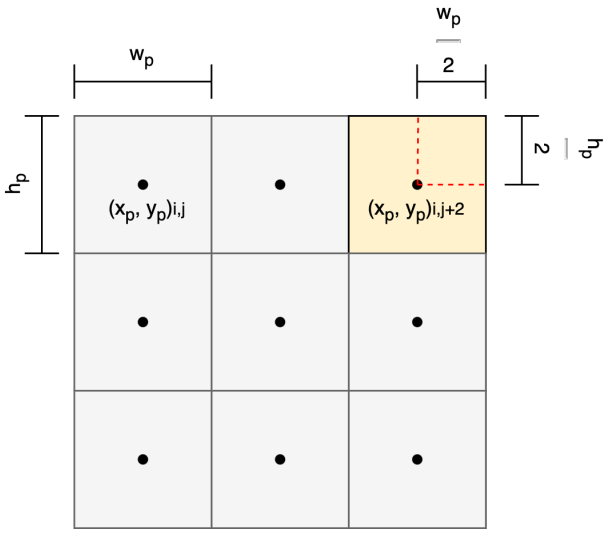

Figure 2. Anchor box generating in the Masking Candidate layer.

Table 1. The statistics of our dataset in each type.

Scene type	Training	Validate	Testing	Source
Landslide	1,690	200	285	BGS and Shipborne
Normal	3,500	669	281	Kaggle

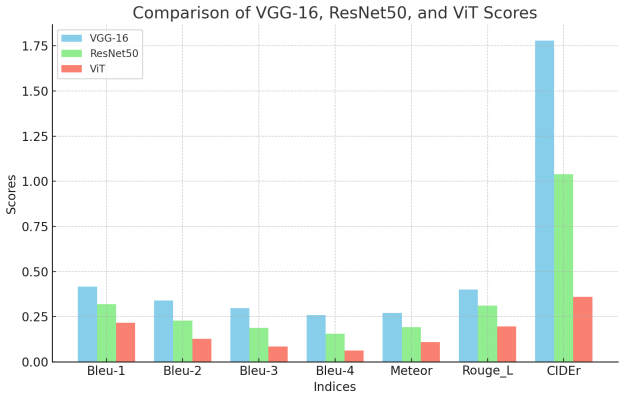

Figure 3. Ablation study of feature extraction models between VGG-16 [10], ResNet50 [7], and ViT [4].

Table 2. Performance metrics for various thresholds in evaluation.

Threshold	Bleu-1	Meteor	Rouge_L	CIDEr
0.2	0.106	0.156	0.252	0.664
0.4	0.156	0.195	0.309	1.075
0.6	0.416	0.270	0.401	1.779
0.8	0.256	0.267	0.395	1.759

phase, while the Shipborne dataset [8] is used in the inference part for testing the performance of DeNet with unseen data purposes.

In non-landslide or normal images, we use the Kaggle image, which contains the natural and environmental scene. We then labeled the Kaggle dataset with the caption in each image. The Kaggle dataset was used in the training phase for the typical scene situation. In the labeled captions, we used simply simple words, such as water, tree, mountain, rocks, soil, people, etc., including the related position of objects. We decided not to use specific words such as 'landslide,' 'mudslide,' etc., to detect the characteristics and related position between objects in the scene.

4. Additional experiment results

In the ablation study, we focus on evaluating three main components of the Detective Network: Feature Extraction, Masking Candidate, and Masking layers.

4.1. Ablation Study on Feature Extraction

We first conducted an ablation study on Feature Extraction, comparing VGG-16 [10], ResNet50 [7], and Vision Transformer (ViT) [4]. In this experiment, we used the evaluation metrics BLUE-1 to BLUE-4, METEOR, ROUGE-L, and CIDEr scores to assess the generated and reference captions for each feature extraction method. The experiment was conducted using the same parameters across all models: seven patch grids, three anchor boxes per patch, 512 output channels, and a consistent loss function.

The results, shown in Figure 3, demonstrate that VGG-16 outperformed ResNet50 and ViT in feature extraction across all indices. This can be attributed to the fact that VGG uses direct encoding, whereas ViT relies on extracting and expanding features to distribute them across attention mechanisms, which may not align well with our architecture and evaluation criteria.

4.2. Ablation Study on Masking Process

The ablation study of vary the threshold in the masking process. In this experiment, we vary the threshold at 0.2, 0.4, 0.6, and 0.8 in Table 2. The result demonstrates that the number of thresholds below half, the indices BLEU-1, METEOR, ROUGE-L, and CIDEr would not capture the objects to generate the caption compared with our reference. However, a suitable threshold of 0.6 could generate the highest evaluation metrics. At the threshold of 0.8, the evaluation metrics have a better score than the lower ones, but they reduce the score compared with the threshold at 0.6, which is represented in Figure 4.

We did the ablation studies of varying the grid patches and anchor boxes to explore the effect of patch or anchor box changes. In Table 3, we set up the experiment to fix the number of anchor boxes at three and vary the grid patches

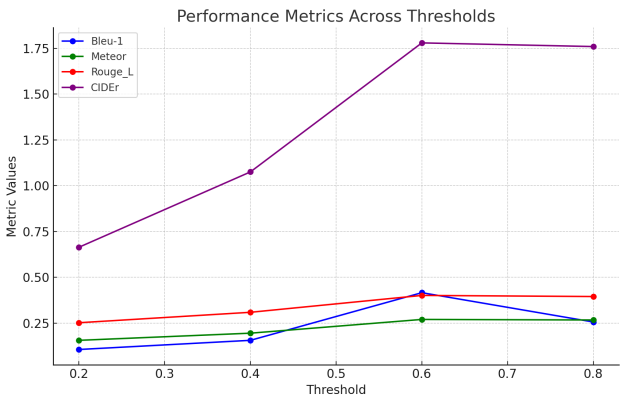

Figure 4. Ablation study of threshold while masking process.

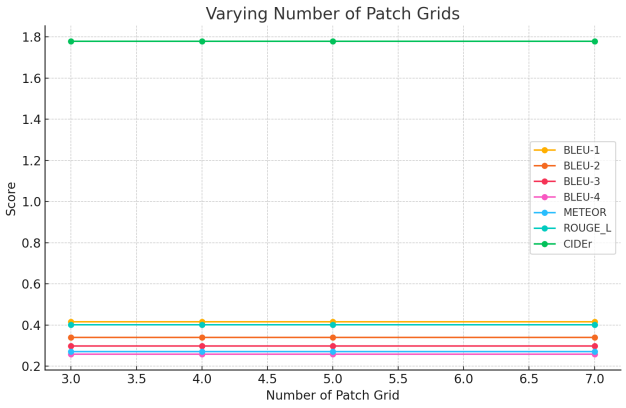

Figure 5. Ablation study of vary patch grids with 3 anchor boxes in each patch.

at 3, 4, 5, and 7 to see the evaluation score of the caption, BLUE-1 to 4, METEOR, ROUGE-L, and CIDEr score. The result of vary grid patches demonstrates in Figure 5 that there is no change in every index of the caption evaluation scores. Therefore, there is no effect on changing the number of patch grids to DeNet performance, which might be due to the Equation 1 to 4 in the Masking Candidate later that adjusts the position and size of a masked region in each patch.

Furthermore, in Table 4, we set up the experiment to fix the number of patch grids at seven while varying the anchor boxes at 3, 5, and 7 to see the evaluation score as in the previous experiment. The result demonstrates in Figure 6 that there is no change in evaluation scores with varying numbers of anchor boxes. In this experiment, we set the anchor boxes at 3 to 5, which might be more than enough for each patch to mask the objects and shift the model's attention. Moreover, in this ablation study, we used seven patch grids with more resolution (49 patches) to cover the major objects. The generated anchor boxes start from $7 \times 7 \times 3$ (3 anchor boxes) to $7 \times 7 \times 5$ (5 anchor boxes), which is enough for cover to shift attention.

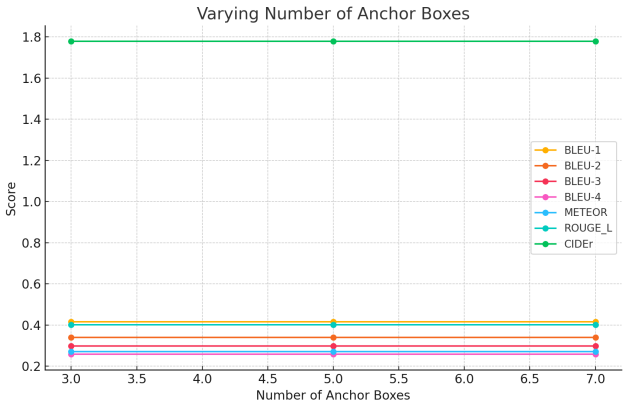

Figure 6. Ablation study of vary anchor boxes with 7 patch grids.

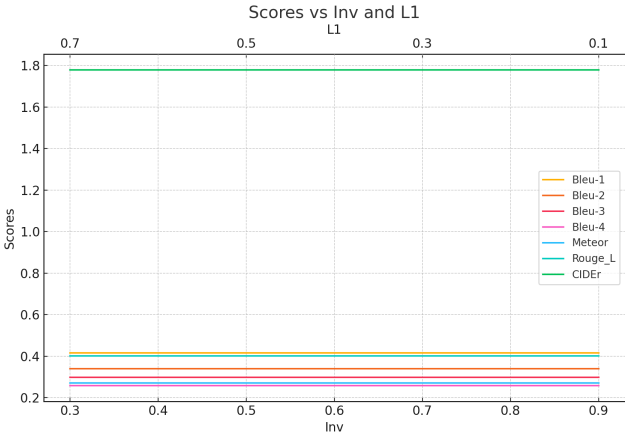

Figure 7. Ablation study of vary Inverse Cosine Similarity loss and L1 Regularization coefficient in each scores.

We did the ablation study of the custom loss function we used from the Inverse Cosine Similarity L_{invc} and L1 Regularization L_1^{reg}. We vary the coefficient of L_{invc} and L_1^{reg} at $(0.9, 0.1)$, $(0.7, 0.3)$, $(0.5, 0.5)$, and $(0.3, 0.7)$. The result shows in Figure 7 that there is no change to the inference of the caption with BLUE-1 to 4, METEOR, ROUGE-L, and CIDEr score.

However, the ablation study of coefficient ration between L_{invc} and L_1^{reg} in the custom loss function demonstrates in Figure 8 that the more coefficient ratio in L_1^{reg}, the more loss we get, and the loss seems more sway which our experiment did in 15 epochs.

4.3. Ablation Study on Masking Layer

We fine-tune the DeNet with fire and flood disasters to evaluate the performance in other disasters. We used the Disaster Dataset (DID) [1] and labeled it with the caption as landslide and normal dataset that we mentioned. Flood and fire images in the training phase used 620 images, while the inference part for testing the performance had 114 images. The performance for a generated caption with DeNet

Table 3. The ablation study of number of patch grid with 3 anchor boxes in DeNet.

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. The ablation study of number of anchor boxes with 7 patch grids in DeNet.

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

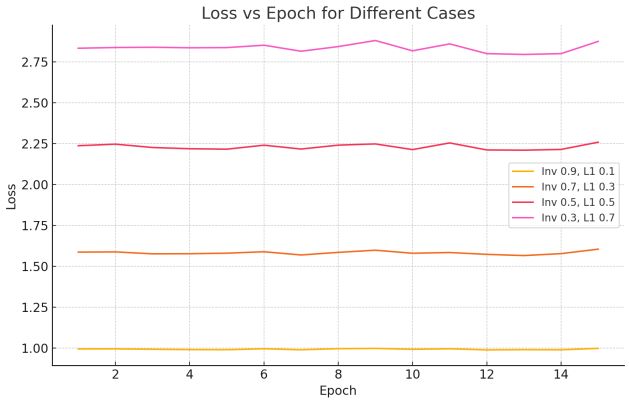

Figure 8. Ablation study of vary Inverse Cosine Similarity loss and L1 Regularization coefficient in epoch.

with BLEU-1, METEOR, ROUGE-L, and CIDEr is demonstrated in Table 5. The result shows that the generated caption compared with the reference caption in flood disaster matches the wording better than in fire disaster; BLEU-1 and ROUGE-L scores of the flood are slightly higher than those fire cases. However, METEOR and CIDEr scores, which consider synonyms and stemming of wording in fire disasters, were slightly higher than in flood cases. Because the flood scenes almost contain only the water body and some construction, which is simple, generated captions will not vary compared with fire disaster scenes.

Figure 9 and 10 show the original image, caption from the original image, masked image after DeNet, caption from the masked image, DeNet attention heat-map, and the target region of the scene of flood and fire disasters. Figure 9 demonstrates the flood disaster in the testing set, which can capture the flood event in the scene quite well. Almost all flood scenes contain only the water body in the significant objects, affecting only a little between captions from VED with DeNet and VED.

Table 5. Performance metrics for landslide, flood, and wildfire disaster in image captioning.

Disaster	BLEU-1	METEOR	ROUGE_L	CIDEr
Flood	0.223	0.146	0.230	0.423
Fire	0.205	0.156	0.219	0.627

Fire disaster scenes in Figure 10 present a more complex challenge due to the presence of multiple objects in the images. Despite this, the generated caption from VED with DeNet manages to improve its performance. The fire event scenes, with their many significant objects, do cause some attention shifting from the first priority. However, DeNet's resilience is evident as it continues to focus on the major object, the fire regions in Figure 10(a).

5. Limitation

The architecture could be improved by increasing the 30-generated text token count. Better information for detec-

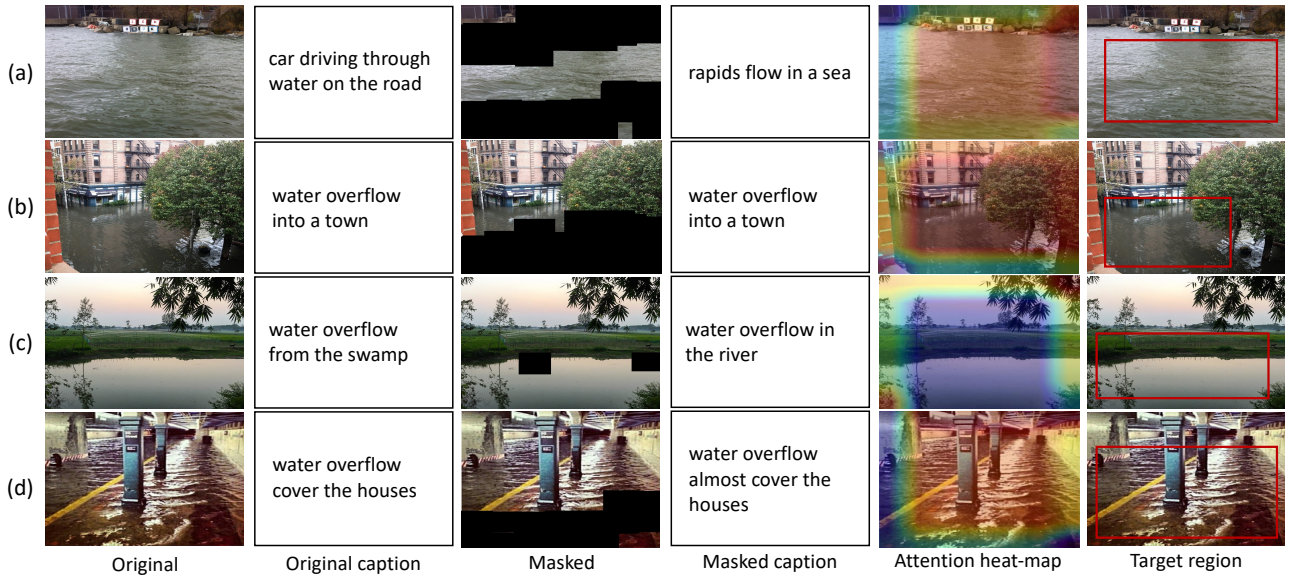

Figure 9. Image captioning results in flood disaster with original image, caption from original image, masked image after DeNet, caption from masked image, DeNet attention heat-map, and target region of scene.

Figure 10. Image captioning results in fire disaster with original image, caption from original image, masked image after DeNet, caption from masked image, DeNet attention heat-map, and target region of scene.

tion could be obtained by improving the caption generation component with a more reliable Natural Language Processing model. Several methods can be used to enhance the Vision Encoder-Decoder's (VED) performance. Utilizing techniques like Shift Windows (SWIN) can aid in sharpening focus on the vision component. Experimenting with various tokenizer types can yield more detailed captions for the caption section.

It could be helpful to implement a looping inference mechanism to shift attention until the target region is identified to address the limitation in detection when there are more than two major objects in the scene, such as Figure 10(a). Furthermore, combining multi-modal and multi-input techniques might yield more thorough data, improving detection accuracy.

References

[1] Disaster image dataset, 2019. 3

[2] Landscape pictures, 2020. 1

[3] The national archive of geological photographs, 2023. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 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, 2014. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[8] Yi Li, Ping Wang, Quanlong Feng, Xiaohui Ji, Dingjian Jin, and Jianhua Gong. 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, 2023. 2

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2