IMAGE CAPTIONING OF ANOMALIES

Beyza Cevik, Harun Alperen Toktas, Mert Cokelek

1 INTRODUCTION

Autonomous systems have gained extreme popularity in various application areas such as self-driving cars, surveillance tasks, autonomous warehouses, and factory systems, and so on. While an autonomous vehicle perceives its environment, it must be prepared for unexpected situations as an abnormality or an outlier behavior in the operation of these systems can be extremely costly. Therefore, classifying and understanding unexpected situations as well as explaining them will assure the cost-efficiency and the quality of the developed systems. In this line of work, one of the most recent research problems on these areas covers determining whether there is an outlier so-called anomaly in the visual content or not. At this point, the undeniable performance of recent deep learning-based computer vision approaches has been motivating the researchers for developing novel models for detecting, classifying, segmenting, or describing the anomalies occurring on images/videos. However, to the best of our knowledge, the proposed approaches so far do not provide descriptive models about general-purpose abnormal cases in visual contents. The main focus of our work can be summarized as: (1) segmenting the anomalous objects and their relations with the environment, (2) generating textual descriptions for these anomalies. Also, we propose to extend the dataset provided in Hendrycks et al. (2019) with detailed captions on the anomalies.

2 RELATED WORK

Studies conducted in the field of anomaly detection vary based on the definition of anomaly, the models used to identify the anomaly behaviour, and usage and application areas. Chalapathy & Chawla (2019) divides anomalies into three types: Point, Collective and Contextual. Point anomaly considered as the outlier data instance that separates from the dataset with high deviation for no reasonable reason. Collective anomaly includes group of data instances that deviates from the dataset and share common characteristics. Contextual anomaly tries to separate regions within data that are contextually irrelevant or out-of-distribution. Recent studies show that deep learning-based models substantially contribute to detection and segmentation of anomalies. In this study, the proposed dataset that includes both collective and contextual anomaly types. The anomalies segmented and textual descriptions are generated by utilizing deep-learning based models. In this study, we are utilizing of supervised deep anomaly detection model besides anomaly segmentation and encoder-decoder image captioning architectures.

2.1 VISUAL ANOMALY DESCRIPTION

Anomaly Detection. This task concerns with the problem of determining whether the input contains outliers or not. The interpretation of outlier is highly dependant on the input data type and the problem. In our line of work, imagevideo anomaly detection problem deals with locating the abnormal objects, scenes and interactions, so-called anomalies. At this point, some of the recent work Hendrycks & Gimpel (2016) based on Deep Neural Networks interpret this problem as a out-of-distribution detection. For this purpose, they perform multi class classification based on the distribution of input data, resulting in a probability distribution of whether being outlier or not. Lee et al. (2017) propose generative-adversarial networks (GANs) to perform out-of-distribution samples by better representing the input data distribution.

Anomaly Classification. Images and videos contain numerous objects and scenes with their relationship and correspondence based on the context. At this point, the classification task can be leveraged and improved on these relationships to determine the type of outliers on the data. The previous approaches on this line of work can be investigated in a broad range from classifying correspondences/actions as anomaly or not as a binary classification Blokus & Krawczyk (2019), further

classifying the anomalies under specific labels such as violence, injury, traffic accidents, disasters, etc. Gayathri et al. (2021).

Anomaly Segmentation. Anomaly segmentation is the task of classifying anomalies in pixel-wise manner on the images. It is a higher-level descriptive problem, compared to the initial works on anomaly detection and classification. In the line of anomaly segmentation, Hendrycks et al. (2019) propose a data set consisting of 7600 images, which are generated by simulated videos, with 1500 anomalous frames, and a total of 13 classes.

2.2 Image Captioning

Describing the visual content is one of the most challenging tasks in artificial intelligence domain since it requires knowledge on both computer vision and natural language processing. At this point, Xu et al. (2015) propose to leverage Convolutional Neural Networks (CNNs) and Long-Short Term Memory (LSTM) networks to generate image captions. This is an encoder-decoder architecture, where the CNN architecture is used for feature extraction on the raw image to further be decoded in the LSTM to generate word by word textual representations.

3 Method

The overview of our proposed two-stage approach is illustrated in Figure 1. In the first stage, the input images are fed to the pre-trained semantic anomaly segmentation model. Later, a linear post-processing is applied on the segmentation results to highlight the anomalous regions to further be fused with the original input images. In the second stage, the fusion results are fed to the image captioning model for fine-tuning, according to the prepared ground truth anomaly captions. Finally, as the output, descriptive anomaly captioning results are expected.

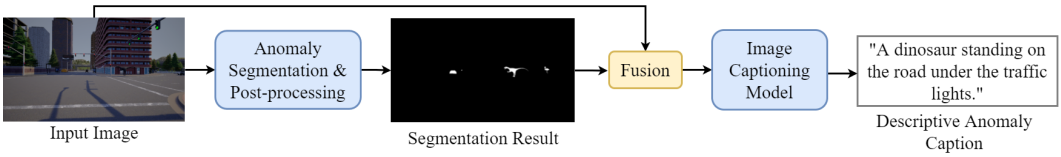

Figure 1: System overview of the proposed approach.

3.1 Anomaly Segmentation Model

There are two widely approaches for image segmentation: (1)Spatial Pyramid Pooling, (2)Encoder-decoder networks. In Hendrycks et al. (2019), encoder-decoder models are used to represent the information. Also, it is shown that using MaxLogit detector to assign anomaly scores performed very well in multi-label Hendrycks et al. (2019) dataset. Therefore, this model is used to obtain segmentation images.

3.2 Post-processing

The Anomalous Object Segmentation with the modifications proposed in Hendrycks et al. (2019) is trained with 12 classes: pedestrian, sidewalk, street lines, road, traffic signs, vehicle, vegetation, wall, building, pole, and fence. Test set which is used as input dataset for captioning model included anomalous objects such as: furniture, animals, military crafts, guns, vehicles etc. These anomaly objects are detected and segmented by the model and any of these objects is considered as the 13rd class. In the post processing step, objects belonging to the 13rd class are masked to further develop an image captioning model by utilizing the information gained from the segmentation model about the anomalous objects.

3.3 FUSION

In order to add a bias on the image captioning operation to produce more descriptive outputs about the anomalies, the obtained anomaly segmentation maps and the input images are proposed to be fused. In the development phase, our aim is to investigate different fusion strategies and their effects on the caption results. In this context, three main fusion strategies considered from low-level to high level have been described as follows.

RGB-A Fusion. Channels of an image provide information obtained from different sensors such as color spaces. An intuitive approach is to fuse the post-processed segmentation image mask which carries descriptor information from the perspective of segmentation. The composition of information from different perspectives may provide a more discriminative and descriptive feature for the image captioning model. At this point, one approach can be expanding the input image channels from 3D to 4D by adding an extra layer that corresponds to the multi-class anomaly segmentation map. The main drawback of this approach is the unavailability of pre-trained models, which are based on 3D input feature space. Hence, in order to leverage the pre-trained image captioning models which have been trained on the large MS-COCO dataset for fine-tuning, we propose to keep the image encoder architecture as is, and add an extra encoder stream for segmented image feature extraction. Later, the output vectors of the two encoders are proposed to be fed to the captioning decoder to produce the final output, which is expected to be more descriptive for anomalies. The 'RGB' stands for Red, Green, Blue and the 'A' stands for Anomaly channels.

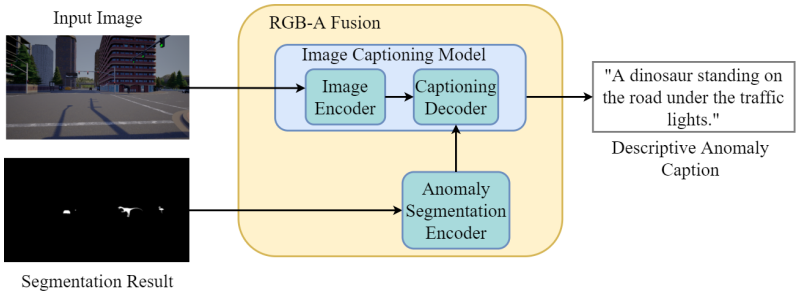

Figure 2: Overview of our proposed RGB-A Fusion approach.

Normalized Late Fusion. Another approach is to use the post-processed segmentation model outputs in addition to the encoder model outputs to achieve an enhanced feature descriptor. You can see the proposed idea in Figure 3.

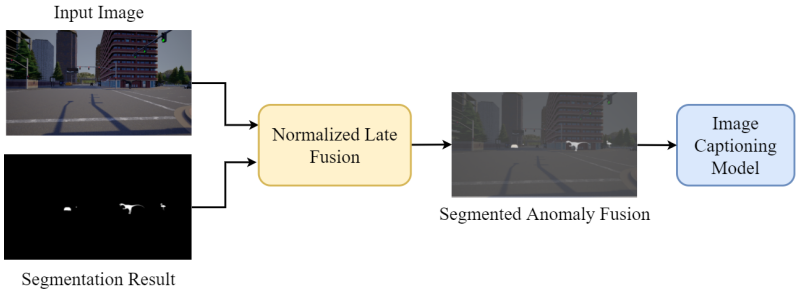

Figure 3: Overview of the Normalized Late Fusion strategy.

Weighting with Segmentation Feature Maps We will scale the feature maps we obtained as a result of the segmentation model within a certain range [0-1], and we will ensure that the attention model gives more weight to the relevant places.

Region of Interest Cropping As you can see in Figure 4, one of the basic approaches to make captioning results tend to describe anomalies is to crop the input images based on the location and size of the contained anomalies to create so-called regions of interest (ROI). The image ROI is then

proposed to be fed into the encoder of the image captioning model for fine-tuning. As a result, the captioning outputs are expected to be more descriptive about the anomalous objects and the relationship with their close peripheries.

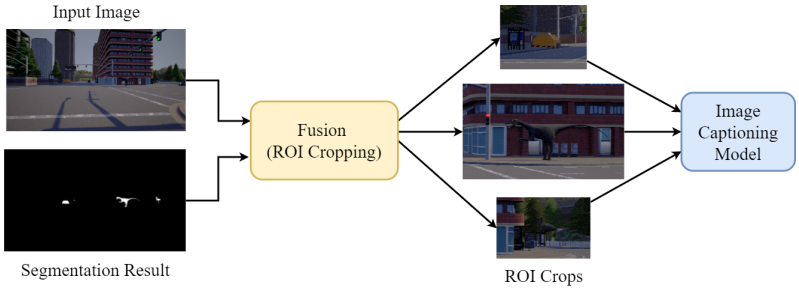

Figure 4: Region of Interest Cropping.

Attention. We can think of the attention mechanism as focusing on the details of the subset in the input image given to learn a task. In our proposed model, we aim to ensure that the areas with anomaly segmentation in the given image create a subset for the relevant input image. In this way, we aim to increase the accuracy of the explanation created by focusing on the correct subset forming the given image.

3.4 Image Captioning Model

To understand what is happening in the images given, the created models should be well represented both semantically and contextually. After making use of the attention mechanism to provide these connections in the given image, we need a structure that will explain these feature maps. The recently used Xu et al. (2015) structure is to feed the extracted feature maps with sequence-to-sequence models such as RNN's and LSTM's. In the model we propose, we will fine-tune the data set we have created on the model we have determined Xu et al. (2015), using the captions we have extracted for the synthetic images created on the Carla simulation.

4 Experimental Settings

4.1 Anomaly Segmentation and Captioning Dataset

Contextual anomaly detection datasets are more difficult to obtain than other detection datasets as we propose to detect and describe the anomaly region in an image. It is due to some difficulties brought by the nature of the problem. Such as (1) anomaly detection datasets suffer from an imbalance in normal and anomaly data instances as there are fewer anomaly data points in a dataset. (2) existing datasets are prepared for only classification tasks. (3) anomaly contexts are restricted to a limited number of contextual contents (surveillance, violent scenes, etc.). The test set of the dataset provided in Hendrycks et al. (2019) for out-of-distribution detection for real-world settings extended with captions.Hendrycks et al. (2019) consists of 7656 image pairs (real-segmentation ground truth) created for out-of-distribution (anomaly) segmentation task. The dataset was split into a train (5125 images), validation (1031 images), test (1500) sets. The train set images includes classes tagged as normal (back-ground, road, street lines, traffic signs, sidewalk, etc.) a while the test set of 1500 images included objects from classes not in the normal classes. Only the test set is extended with captions to achieve a dataset compatible with anomaly segmentation and captioning models. The dataset contains images synthetically produced by the Carla simulation tool and Unreal Engine using different angles, distances, lights, and colors in two different town simulations to obtain realistic images. Images contain unexpectedly large-big objects such that huge sofas, drum kits, digital circuits, butterflies, grenades, military vehicles in a total of 250 kinds of out-of-distribution regions, and objects. These objects are generally huge objects that we do not expect to be located in the city or appear in unexpected places when viewed at any point in the city. We will make use of 1000 images of the data set consisting of 1500 images in total. 1000 images from the dataset were

captioned by all three authors and will achieved 3k captions. Currently we finished half of them. Sample images in the dataset used have been provided in Figure 5.

4.2 Development Environment

We also set up a system on Google Colab Çevik (2021) to explain images easily; The images come in order, as we type, we move on to the next and whenever we want, we continue from where we left off. We've now completed half of the total captions that we will type. We run all the code work on Google Colab. We use our Google Colab Pro accounts to train the models. We carry out model training using TeslaP100 GPU's.

4.3 Evaluation

There are three metrics to evaluate in the segmentation section. These are AUROC, FPR95 and AUPR metrics. We will not make a progress in the segmentation part, but we can still share the results on the existing data set. We use the Bleu score to evaluate the results generated by Image Captioning part. It simply compare the produced caption and the caption given as reference(label).

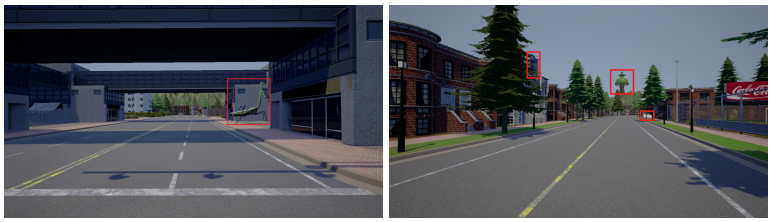

Figure 5: Sample anomalous images in the dataset.

5 Experimental Results

We have performed different training modalities which are covered in the previous section.

5.1 Training with Region of Interest Cropping

In this approach, we have tried to crop the anomalous objects and their peripheries, and feed those regions to the captioning model, instead of the entire image. Our motivation behind this approach was to force the generated captions to focus on the anomalous objects, and their relationships with their peripheries. However, in the cases where there are multiple anomalous objects on a single frame, this approach fails. It's because the ground truth captions are created such that all anomalies are mentioned in a single sentence. After creating multiple crops per anomalies, the one-to-one relationship between anomaly crops and the ground truth captions are missed.

5.2 Training with Normalized Late Fusion Strategy

In another experiment, we have trained the encoder and the decoder parts of the captioning model together. As mentioned in the original paper, the initial learning rates for the encoder and the decoder are selected as 1e-4 and 4e-4, respectively, with ADAM optimizer. The batch sizes are selected as 32 and 60 for the encoder and the decoder. The model is trained with the Cross Entropy Loss for 20 epochs. The training procedure is summarized in the accuracy-loss-BLEU score curves in the figures below.

The training curves look promising, however, the validation process was not that successful. We believe that the model has overfitted. The reason of overfitting is that the train and validation sets are split such that validation set contains anomalous objects that do not occur in the train set. And its main reason is the small size of the dataset. Since we can not increase the amount and the variaty of the dataset, in the rest of the experiments we divided it into train and test splits not video-wise but frame-wise. In this way, both of the splits contain common anomalous objects, but with different viewing angles, distances, and lightning conditions.

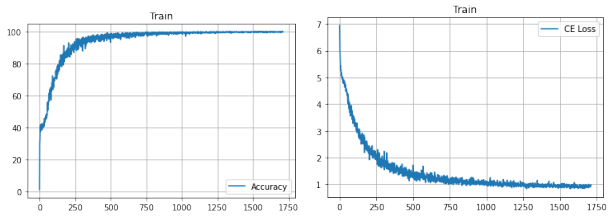

Figure 6: Train set accuracy and CE Loss curves of Normalized Late Fusion Training.

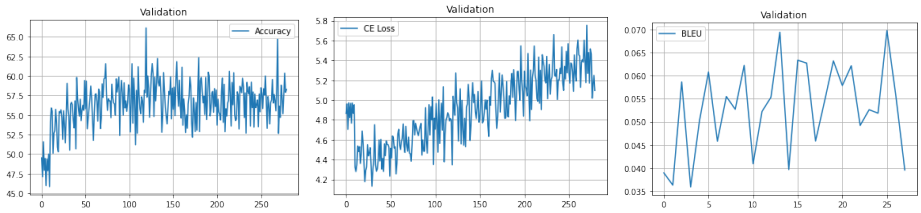

Figure 7: Validation set accuracy, CE Loss and BLEU Score curves of Normalized Late Fusion Training.

5.3 Training with RGB-A Fusion

In our first experiment, we have applied a two-stream strategy. The first branch contains an encoder, based on ResNet-101 feature extractor which outputs feature representations of the raw RGB input image. Similarly, the second branch has the same encoder architecture, which independantly extracts features from the masked anomaly images. The input dimension of the encoders is (3 x 256 x 256), hence the images are resized to 256x256 in the pre-processing step. And the output dimension of the encoders is (14 x 14 x 2048), which is the same as the output of the ResNet-101 feature extractor. The outputs of the encoders are then concatenated with each other and applied 2048 1x1 convolutions so that the final feature matrix - also the input of the captioning decoder- is obtained with a size of (14 x 14 x 2048). This matrix is then fed to the LSTM-based decoder to produce the caption predictions.

The first stream of the model is trained with pre-trained weights of the ResNet-101-based encoder, and the other stream is randomly initialized, since the input space (masked anomaly images) are different than the onr of ImageNet dataset. The decoder is also randomly initialized. The training is performed with a batch size of 32, initial learning rates from 1e-3 to 4e-6 and ADAM optimizer for at most 20 epochs. Also, early stopping based on the BLEU score is applied. The accuracy, Cross Entropy Loss, BLEU Score curves for the training and validation sets are given in the following figures.

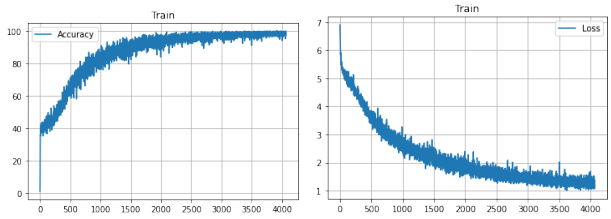

Figure 8: Train set accuracy & CE Loss curves of RGB-A Fusion Training.

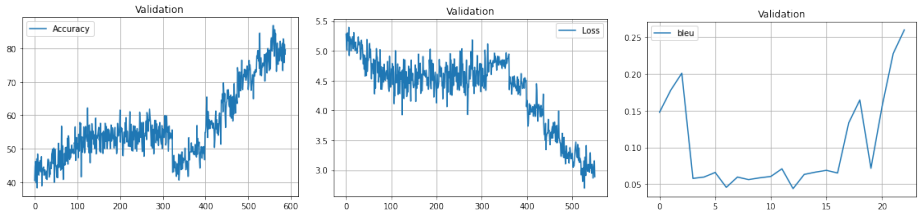

Figure 9: Validation set accuracy, CE Loss and BLEU Score curves of RGB-A Fusion Training.

6 Conclusion

In this project, we have addressed the problem of generating descriptive captions for anomalies. We have proposed a total of 3000 descriptive ground truth captions for 1000 video frames, for an existing dataset -Street Hazards-. We have investigated the performance of different encoding-decoding strategies for anomaly feature extraction, fusion and captioning such as RoI Cropping, Normalized Late Fusion and RGB-A Fusion. The qualitative and quantitative results demonstrate that the RGB-A Fusion can be a good turning point for anomaly captioning, however the low dataset size caused the training process not to complete successfully. Increasing the amount of dataset and developing novel architectures is left as future work.

References

Adam Blokus and Henryk Krawczyk. Systematic approach to binary classification of images in video streams using shifting time windows. *Signal, Image and Video Processing*, 13(2):341–348, 2019.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *CoRR*, abs/1901.03407, 2019. URL `http://arxiv.org/abs/1901.03407`.

R. G. Gayathri, Atul Sajjanhar, Yong Xiang, and Xingjun Ma. Multi-class classification based anomaly detection of insider activities. *CoRR*, abs/2102.07277, 2021. URL `https://arxiv.org/abs/2102.07277`.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Beyza Çevik. Image captioning automatization, 2021. URL `https://colab.research.google.com/drive/1-QW00xgXckLvlmUDzIR7Q05mjvH0L8n5?usp=sharing`.

APPENDIX

A Initial Results on Image Captioning

Firstly, we looked at what the Image Caption model Xu et al. (2015) produced on the dataset we used. When we directly give the image, the caption model did not produce any good results regarding the anomaly region, since the regions we call anomaly are relatively small compared to entire image. When patch the anomaly

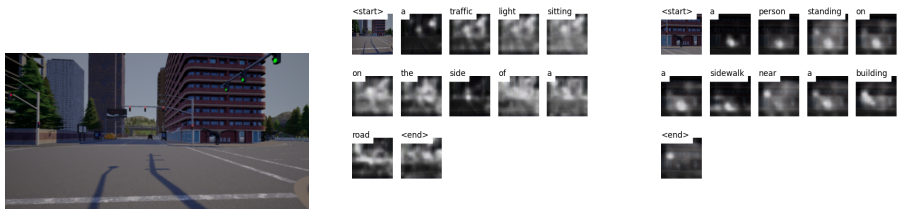

Figure 10: Initial captioning outputs of image & segmented anomaly pairs

regions instead of giving the whole picture, the results produced became more specific. This shows that the system we designed can produce more detailed results about anomaly regions.

B Sample Prepared Ground Truth Captions

Figure 11: A sample image on the dataset and the corresponding captions prepared by us: (1) trees flying on the mountain and a rocket launcher behind the fences (2) there is a wheel in the bushes and there are military aircraft on the hill and in the trees. (3) there is a green tool on the top of the mountain

C Final Results on Image Captioning

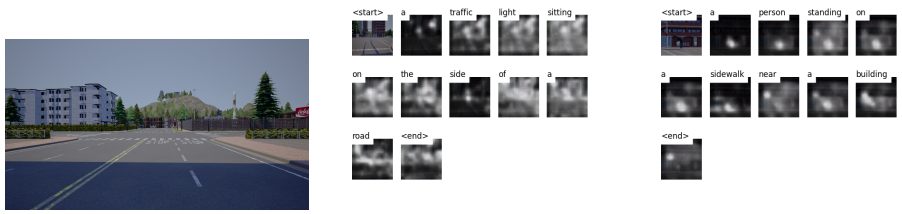

Figure 12: Final captioning outputs of image & segmented anomaly pairs, obtained by training with the RGB-A Fusion