EDTER: Edge Detection with Transformer

Liu Weihao
A0232935A
liuweihao@u.nus.edu

1. Summary

1. What is the main problem the paper tries to solve?

   In this research paper, the author's primary objective is to enhance the efficacy of edge detection techniques by incorporating transformer-based models into the process [3]. Despite the potential advantages of utilizing transformers in edge detection, there are two significant challenges that must be addressed to ensure optimal performance:

   I) Owing to computational constraints, transformers are frequently employed on relatively large-sized patches. However, these coarse-grained patches are not conducive to learning accurate features for edges. Therefore, it is imperative to implement self-attention mechanisms on fine-grained patches without exacerbating the computational demands of the system.

   II) The task of accurately extracting edges from intersected and thin objects poses a considerable challenge in the field of edge detection. To address this issue, the authors investigate novel methodologies for the identification and extraction of precise edges, even in the presence of complex, intersected, and thin objects, thus significantly improving the performance of edge detection algorithms.

2. Why is the problem important theoretically?

   Edge detection serves as a critical cornerstone in the realm of computer vision, providing the foundation for numerous theoretical advancements and innovations in the field [2]. The primary goal of edge detection is to accurately identify and extract the boundaries of objects and visually prominent edges within a given input image, thereby facilitating the accurate representation and interpretation of visual information in both static and dynamic scenes. As a fundamental problem in computer vision, it has led to the development of various techniques and models that push the boundaries of performance and accuracy.

   In recent years, transformers have emerged as a powerful tool for various tasks in natural language processing and computer vision. Transformers offer a unique approach to processing data, utilizing self-attention mechanisms to capture long-range dependencies and complex contextual information. The incorporation of transformer-based models into edge detection processes has the potential to significantly enhance the performance of edge detection algorithms by enabling better representation and understanding of visual features, particularly in challenging scenarios with intricate backgrounds, inconsistent annotations, varying illumination conditions, and diverse object textures.

   Applying transformers to edge detection brings a new level of theoretical importance to the field, as it offers the opportunity to overcome some of the inherent limitations of traditional edge detection techniques. For instance, by addressing challenges like processing fine-grained patches without increasing computational burden and extracting precise edges from intersected and thin objects, transformers can contribute to the development of more advanced edge detection algorithms that are both computationally efficient and highly accurate. Furthermore, the successful integration of transformers into edge detection methodologies can serve as a stepping stone for the development of more sophisticated computer vision systems, capable of tackling a wider range of problems and tasks. Ultimately, the theoretical advancements resulting from the fusion of edge detection and transformer-based models will have far-reaching implications, driving innovation and progress in the broader field of computer vision.

3. Why is the problem important in practice? What are the applications?

   The practical importance of edge detection lies in its broad applicability across a wide range of real-world scenarios. Applications such as image segmentation, object detection, video object segmentation, facial recognition, and augmented reality all rely on effective edge detection as an essential component of their underlying processes. Some notable examples of

edge detection in practice include autonomous navigation, where vehicles must perceive and understand their surroundings to navigate safely; medical imaging, where accurate edge detection can aid in the diagnosis and treatment planning of various conditions; and visual content analysis, employed in diverse industries for purposes like video surveillance, quality control, and media production.

Moreover, the incorporation of advanced techniques such as transformer-based models into edge detection processes further enhances the practical utility of this fundamental computer vision concept. By addressing challenges like processing fine-grained patches without increasing computational burden and extracting precise edges from intersected and thin objects, these innovations drive improvements in the performance and efficiency of edge detection algorithms. This, in turn, contributes to the development of more robust and reliable computer vision systems that can perform effectively across a wide range of practical applications, unlocking new possibilities and enabling further innovation across various domains.

4. What are the key ideas of the paper that differentiate it from other papers/methods?

   In order to address the first challenge, this research paper incorporates transformer-based models into the edge detection process. To tackle the second challenge, a two-stage approach is designed, with the aim of exploring long-range global context (Stage I) and capturing fine-grained local cues (Stage II). Additionally, a Feature Fusion Module (FFM) is employed to effectively combine the cues extracted from both Stage I and Stage II, resulting in a more comprehensive and accurate representation of the visual features within the edge detection framework.

5. What are the contributions/novelty made by the paper?

   I) The authors propose a novel transformer-based edge detector, Edge Detection TransformER (EDTER), to detect object contours and meaningful edges in natural images. It is the first transformer-based edge detection model.

   II) EDTER is designed to effectively explore long-range global context (Stage I) and capture fine-grained local cues (Stage II). Moreover, the authors propose a novel Bi-directional Multi-Level Aggregation (BiMLA) decoder to boost the information flow in the transformer.

   III) To effectively integrate the global and local information, the authors use a Feature Fusion Module (FFM) to fuse the cues extracted from Stage I and Stage II.

   IV) Extensive experiments demonstrate the superiority of EDTER over the state-of-the-art methods on three well-known edge detection benchmarks.

6. State whether each of the contributions is significant. Justify your statement (again, for each claimed contribution/novelty).

   I) The significance of this contribution lies in the adoption of transformers, which have emerged as powerful tools for various tasks in both natural language processing and computer vision domains. Transformers provide a unique approach to data processing, utilizing self-attention mechanisms to capture long-range dependencies and complex contextual information, which is crucial for enhancing the performance of edge detection algorithms.

   II) The importance of this contribution is also evident in the innovative two-stage training approach, which effectively improves the model's performance. Moreover, the BiMLA decoder accelerates the flow of information, further optimizing the overall process.

   III) This contribution is vital in its ability to seamlessly integrate the outcomes of Stage I and Stage II. By effectively combining the information from both stages, the model is able to generate a more comprehensive and accurate representation of the visual features within the edge detection framework.

   IV) The importance of this contribution extends to its versatility, as demonstrated by the successful application of the proposed method across various datasets. This adaptability showcases the potential of the method to be employed in diverse contexts, thereby broadening its scope and impact on the field of computer vision.

7. How is the performance compared to the SOTA methods? Significantly better, marginally better? Justify your answer.

   I) With regard to the BiMLA decoder and the Feature Fusion Module (FFM), experimental results demonstrate that the proposed method significantly outperforms SOTA approaches in edge detection. These findings indicate the effectiveness of the BiMLA decoder and FFM in enhancing the accuracy and precision of edge detection algorithms, thereby contributing to the advancement of the field.

   II) The proposed method achieves superior results across all evaluated datasets, showcasing its robustness and adaptability. The EDTER model marginally surpasses the performance of state-of-the-art methods, exhibiting improvements in the range of 0.02-0.03 in

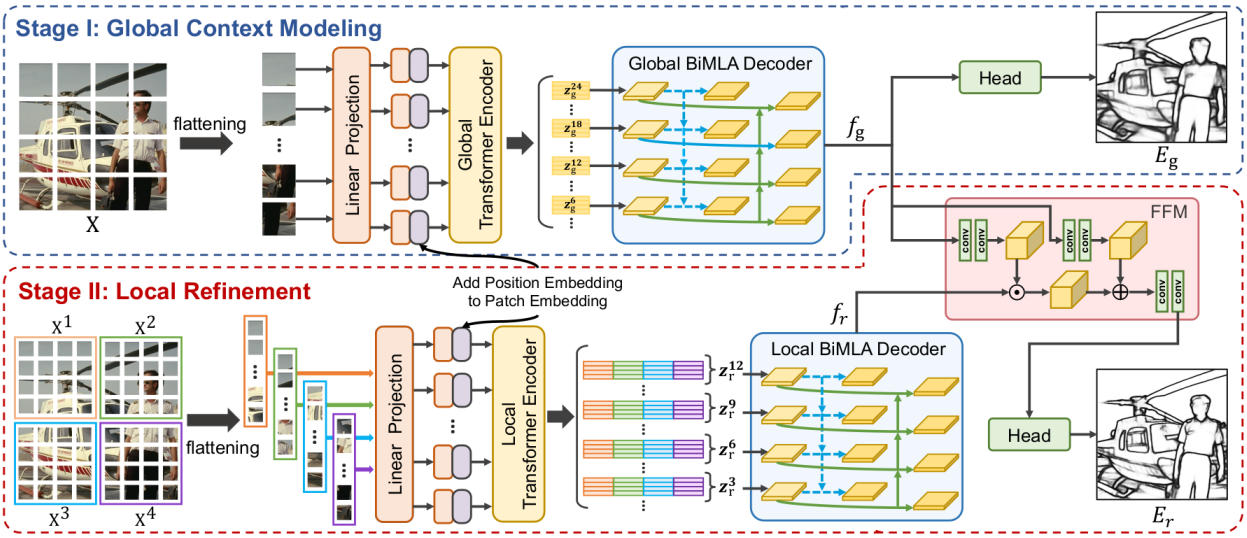

Figure 1. Overall framework

performance metrics for the three types of datasets analyzed. This consistent enhancement of edge detection performance demonstrates the potential of the proposed method to be successfully employed in diverse scenarios, further solidifying its value as a significant contribution to the domain of computer vision.

8. Summarize the algorithm proposed in the paper by showing the proposed pipeline or architecture, whenever applicable.

   As shown in Figure 1, the proposed model consists of two primary components. Stage I focuses on Global Context Modeling, wherein a global transformer encoder calculates the global attentions. Following this step, a global BiMLA decoder generates high-resolution features, which are then used to predict edge maps via a decision head. In Stage II, a process similar to Stage I is employed, with partitioned patches being inputted into a local transformer encoder to produce local attentions. The concatenated attentions from both global and local contexts are harnessed to decode the high-resolution features. Finally, a decision head predicts the edge maps using the features derived from both Stage I and Stage II, which are effectively fused by the Feature Fusion Module (FFM). This comprehensive approach enables the model to capture and integrate diverse contextual information, ultimately contributing to the enhanced performance of the edge detection algorithm.

9. If you have anything else to say, write here:

2. Experiment and Analysis

1. How did you do your experiments?

   The authors have made their source code publicly available, allowing for independent replication and evaluation of their proposed method. I utilized the provided code to train the model on the BSDS500 dataset, incorporating both Stage I and Stage II in the training process. Upon obtaining the trained model, I proceeded to test its performance on the NYUDv2, Multicue, and COCO datasets [1]. Notably, the COCO dataset represents a novel addition to the evaluation, as it was not explicitly mentioned in the original research paper. However, there are 5k images in COCO test dataset, I randomly select 200 images and input them into the trained model.

2. What are the reasons for choosing your own data?

   The choice of the COCO dataset as an additional test dataset for edge detection is driven by its unique characteristics that make it particularly suitable for evaluating the proposed method's performance:

   (a) Complex Scenes: The COCO dataset consists of images with complex scenes, containing multiple overlapping objects and varying levels of detail. This complexity allows for a more robust assessment of the edge detection method's ability to handle intricate scenarios and accurately identify object boundaries.

   (b) Varied Object Sizes: COCO images include objects of different scales, from large, dominant objects to smaller, less conspicuous ones. Eval-

uating the model on this dataset tests its ability to detect edges in objects of varying sizes, which is crucial for a comprehensive edge detection method.

(c) Rich Object Annotations: The COCO dataset provides detailed object annotations, including segmentation masks and object categories. This rich annotation data enables a more precise evaluation of the edge detection results, allowing for a better understanding of the model's performance in terms of both localization and classification accuracy.

(d) Diverse Object Categories: COCO features 80 different object categories, encompassing a wide range of object types, from animals and vehicles to furniture and electronic devices. This diversity allows the model to be tested on a broad set of object classes, providing insights into its ability to perform edge detection across various domains.

3. What are the experiment results of using the data used in the paper? Show them here. The result shown here must be generated by yourself, and not copied from the paper.

   In order to evaluate the performance of the proposed Edge Detection Transformer (EDTER) model, I initially adhered to the methodology outlined in the paper, training the model on the BSDS500 dataset and subsequently using its test dataset for verification purposes. As depicted in Fig 2, the Stage I model is capable of obtaining a general representation of the overall boundaries but lacks precise accuracy. In contrast, the Stage II model effectively extracts clear boundaries and edges by leveraging both global and local cues, demonstrating the improved performance of the proposed method in capturing and integrating diverse contextual information for edge detection tasks.

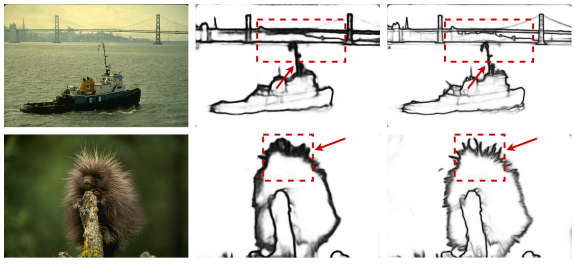

Figure 2. Train the model on BSDS500 train dataset, and test on BSDS500 test dataset. From left to right are the input images, the results of the Stage I model and Stage II model, respectively.

To assess the robustness of the proposed model, I applied the model trained on the BSDS500 dataset to the NYUDv2 and Multicue datasets, as mentioned in the research paper. The outcomes are presented in Fig 3 and Fig 4. It is noteworthy that, despite not being trained on the same dataset, the model is still capable of identifying clear boundaries when utilizing the Stage II model, demonstrating its adaptability and generalization capabilities across diverse datasets.

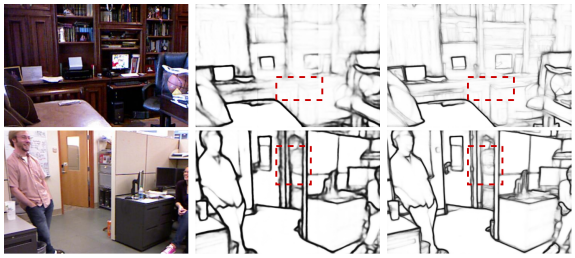

Figure 3. Train the model on BSDS500 train dataset, and test on NYUDv2 test dataset. From left to right are the input images, the results of the Stage I model and Stage II model, respectively.

Nonetheless, as demonstrated in Fig 4, the application of the Stage I model results in dividing lines appearing at the 1/3 and 2/3 positions within the image. Interestingly, when the Stage II model is employed, these dividing lines effectively vanish, indicating the enhanced performance and capability of the Stage II model in refining edge detection outcomes.

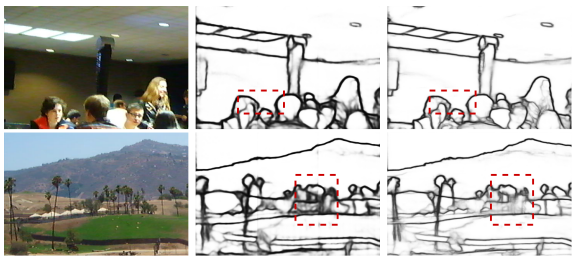

Figure 4. Train the model on BSDS500 train dataset, and test on Multicue test dataset. From left to right are the input images, the results of the Stage I model and Stage II model, respectively.

4. What are the experiment results of data you curated yourself? Show the comparisons? Show them here.

   Following a similar approach to the one employed in item 3, I applied the Stage I and Stage II models, trained on the BSDS500 dataset, to the COCO test dataset. This dataset contains a number of complex scenarios, as exemplified in Fig 5, where a crowded street scene is depicted. The Stage I model is capable of capturing global features and delineating the boundaries of some prominent targets. However, as observed in the Multicue dataset, two dividing lines appear once

again. Upon applying the Stage II model, more intricate details and accurate boundaries are obtained, demonstrating the model's enhanced performance in handling complex and diverse edge detection scenarios.

Figure 5. Train the model on BSDS500 train dataset, and test on COCO test dataset. From left to right are the input images, the results of the Stage I model and Stage II model, respectively.

5. Based on your own experiments, how do you analyse the proposed method?

   The proposed model employs a two-stage training approach to capture both long-range global context and short-range local cues. Subsequently, the Feature Fusion Module (FFM) is utilized to integrate these two models. As illustrated in Fig 5, the Stage I model effectively captures the majority of the global information, particularly in complex scenarios, where this information is typically located in the foreground of the image. Upon the completion of Stage II and the application of FFM, the global information is retained, while additional details are incorporated. Consequently, the boundaries become significantly more accurate, showcasing the effectiveness of the proposed method in enhancing edge detection performance.

6. Based on your own experiments and analysis, what are your conclusions?

   Based on the experimental findings and model analysis, it can be concluded that the Edge Detection Transformer (EDTER) presents a robust and effective solution for edge detection tasks. The two-stage approach, comprising Stage I for global context modeling and Stage II for capturing fine-grained local cues, allows for the extraction of accurate and clear boundaries across diverse and complex scenarios.

   The model's performance on the BSDS500, NYUDv2, Multicue, and COCO datasets demonstrates its adaptability and generalization capabilities, successfully handling various types of images and edge detection challenges. Moreover, the Stage II model's ability to refine the results obtained from the Stage I model, mitigating issues such as dividing lines and enhancing edge accuracy, highlights the effectiveness of the proposed method.

7. If you have anything else to say, write here:

3. Strengths

1. Is the problem important? If yes, why? If not, ignore this question.

   Addressing the two main challenges associated with edge detection is vital for enhancing the performance and applicability of computer vision models:

   (a) The first challenge revolves around the application of transformers to relatively large patches due to computational constraints. Coarse-grained patches are not ideal for learning accurate features for edges, which makes it crucial to perform self-attention on fine-grained patches without increasing the computational burden. By resolving this issue, the efficiency and effectiveness of edge detection models can be improved, enabling them to handle higher-resolution images and intricate details with greater precision. This, in turn, enhances the overall performance of computer vision tasks, such as object recognition, segmentation, and tracking, where accurate edge detection plays a pivotal role in achieving reliable results.

   (b) The second challenge concerns the extraction of precise edges from intersected and thin objects. Accurate edge detection in such scenarios is essential for a wide range of applications, including medical imaging, where discerning the boundaries between overlapping structures is critical for accurate diagnosis and treatment planning. Similarly, in the context of autonomous vehicles and robotics, the ability to detect and distinguish fine, intersecting edges can significantly improve navigation, obstacle avoidance, and object manipulation capabilities.

2. Is the main idea to solve the open problem sufficiently novel? If yes, why? If not, ignore this question.

   Yes, it's novel. The EDTER offers a sufficiently novel solution to the aforementioned challenges, showcasing a unique and innovative approach to edge detection that effectively addresses the issues of transformer-based patch sizes and precise edge extraction from intersected and thin objects. The authors first use transformer on edge detection.

3. Does the main idea make sense and address the open problem? If yes, why? If not, ignore this question.

Yes, the EDTER model introduces a two-stage training method that captures both long-range global context and short-range local cues, thereby improving edge detection performance. This approach enables the model to perform self-attention on fine-grained patches without increasing computational complexity, overcoming the limitations associated with applying transformers to larger patches. The inclusion of the Feature Fusion Module (FFM) further enhances the model's ability to integrate the information extracted from both stages, resulting in more accurate and detailed edge detection.

In addition, the EDTER model is particularly adept at extracting precise edges from intersected and thin objects. The two-stage training process, which leverages global and local cues, allows for the accurate delineation of object boundaries even in challenging scenarios involving overlapping and fine structures. This capability is essential for various real-world applications, where the ability to detect and distinguish fine, intersecting edges can significantly impact overall performance and reliability.

4. Is the idea impactful? If yes: Why? What is the degree of the impact? Explain further how it will impact the related work or downstream applications. If the idea is not impactful, ignore this question.

   The Edge Detection Transformer (EDTER) is an impactful solution in the field of computer vision due to its innovative approach to edge detection and its ability to address critical challenges effectively

   (a) The two-stage training method is capable of capturing both long-range global context and short-range local cues, resulting in a more comprehensive and accurate representation of object boundaries. This approach enables the model to overcome limitations associated with traditional transformer-based edge detection methods.

   (b) The incorporation of the FFM further enhances the EDTER model's effectiveness by facilitating the seamless integration of information extracted from both stages, leading to more accurate and detailed edge detection results. Moreover, the model leverages the Bi-Directional Multi-Level Aggregation (BiMLA) decoder to accelerate information flow, further contributing to the model's efficiency and adaptability across various datasets.

5. Are the ablation studies properly done? If yes, why? If not, ignore this question.

   Yes, the authors give the impact of all the components.

   (a) They analyse the impact of the BiMLA decoder by comparing it with the SETR-MLA decoder.

   (b) They also try to remove the FFM from Stage II to observe the impact of FFM.

6. What are the strengths of this paper according to the paper itself? For each of the claimed strengths, do you agree? Provide your justification.

   (a) It is the first transformer-based edge detection model.

   (b) The authors design a two-stage training approach to explore the long-range global context and fine-grained local cues at the same time.

   (c) To effectively integrate the global and local information, the authors use a Feature Fusion Module (FFM) to fuse the cues extracted from Stage I and Stage II.

7. Based on your experiment and analysis in the previous section, what are the strengths of the paper beyond the claimed strengths in the previous question? Provide your justification. Your justification must be based on your experiment shown in the previous section by providing specific links to the specific experiments shown in the previous section.

   The demonstrated robustness of this model is highly impressive. Remarkably, when trained on a single dataset, the model consistently exhibits outstanding performance across various other datasets, showcasing its ability to generalize effectively and maintain high accuracy in diverse scenarios.

8. If you have anything else to say, write here:

4. Weaknesses

1. Is the problem important? If not, why? If yes, ignore this question.

2. Is the main idea to solve the open problem sufficiently novel? If not, why? If yes, ignore this question.

3. Does the main idea make sense and address the open problem? If not, why? If yes, ignore this question.

4. Is the idea impactful? If not, why? If the idea is indeed impactful, ignore this question.

5. Are the ablation studies properly done? If not, why? If yes, ignore this question.

6. What are the weaknesses of this paper theoretically? Provide your justification for each theoretical weakness.

7. Based on your experiment and analysis in the previous section, what are the weaknesses of the paper? Provide your justification for each stated weakness. Your justification must be based on your experiment shown in the previous section by providing specific links to the specific experiments shown in the previous section.

   According to the paper, the model obtained from Stage I is capable of being applied to images of any size, a finding that was corroborated during my experiments. However, the Stage II model initially divides the image based on size. Consequently, if we wish to apply the model to different images, it is necessary to first crop the image to ensure compatibility with the trained model, allowing for proper data handling and processing.

8. If you have anything else to say, write here:

5. Possible Improvements

1. Based on your statement on the theoretical weaknesses, how can you address each of the weaknesses?

2. Based on your statement on the weaknesses that you found empirically, how can you address each of the weaknesses?

   Addressing the aforementioned issue of applying the Stage II model to images of varying sizes can be achieved through a preprocessing step involving padding or cropping the images. By manually adding padding to the images, we can ensure that they conform to the required size while retaining the original content. Alternatively, cropping the images allows us to focus on specific regions of interest while maintaining compatibility with the trained model. Implementing either of these preprocessing techniques ensures that the input images meet the size constraints of the Stage II model, enabling proper data handling and processing while still capitalizing on the model's ability to extract accurate features and perform effective edge detection. This approach not only enhances the model's applicability across diverse scenarios but also demonstrates its flexibility and adaptability in handling images of various sizes and complexities.

3. If you implement your own ideas to address the weaknesses, describe your algorithm and show your results.

4. If you have anything else to say, write here:

References

[1] Coco (microsoft common objects in context), 2014. 3

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[3] Mengyang Pu, Yaping Huang, Yuming Liu, Qingji Guan, and Haibin Ling. Edter: Edge detection with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1402–1412, 2022. 1