← Back to **Author Console** (/group?id=thecvf.com/CVPR/2024/Conference/Authors#your-submissions)

Co-Attention Bottleneck: Explainable and Causal Attention Emerged from Transformers Trained to Detect Images Changes

(/pdf?id=EdmoFBoWVz)

Pooyan Rahmanzadehgervi (/profile?id=~Pooyan_Rahmanzadehgervi1), Hung Huy Nguyen (/profile?email=hhn0008%40auburn.edu), Peijie Chen (/profile?id=~Peijie_Chen2), Long Mai (/profile?id=~Long_Mai2), Anh Nguyen (/profile?id=~Anh_Nguyen1)

01 Nov 2023 (modified: 26 Nov 2023) CVPR 2024 Conference Submission Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=EdmoFBoWVz)

Authors Confirmed: I understand and agree to the following: After the registration deadline (Nov. 3), authors cannot be added or deleted, only the order can be changed. All authors are *required* to have an up to date OpenReview profile by the paper submission deadline, Nov. 17. Authors who are added to a submission via specifying a name and email address need to create a new OpenReview profile. All author profiles should be updated to include: recent email addresses, career positions, and publications; see https://cvpr.thecvf.com/Conferences/2024/OpenReviewAuthorInstructions (https://cvpr.thecvf.com/Conferences/2024/OpenReviewAuthorInstructions). Papers with one or more authors without an updated OpenReview profile by Nov. 17 may be desk-rejected.
Student Paper: Yes

Abstract:
Vision Transformers (ViTs) are popular but their attention maps are not self-explainable and require a post-hoc feature attribution method to aggregate into a single heatmap. In this work, we propose to use a 1-head, Co-Attention layer at the end of a Transformer encoder (before the final classification head) to serve as an exact, explainable, and editable attention bottleneck. We find that when trained to predict whether there is an object-level change between two images (\ie image difference prediction), the co-attention bottleneck of our binary classifier can remarkably spot the differences far better than existing weakly-supervised ViT models on this task and comparable to change detectors trained explicitly with bounding-box supervision. Furthermore, our co-attention bottleneck is editable and shows a strong causal relationship with the classifier's prediction---a property that is shown for the first time in the literature.

Closest Subject Area That Your Submission Falls Into: Explainable computer vision
Guidelines Confirmed: I confirm that I checked and agree to the author (https://cvpr.thecvf.com/Conferences/2024/AuthorGuidelines (https://cvpr.thecvf.com/Conferences/2024/AuthorGuidelines)) and ethics guidelines (https://cvpr.thecvf.com/Conferences/2024/EthicsGuidelines (https://cvpr.thecvf.com/Conferences/2024/EthicsGuidelines)).
Supplementary Material: pdf (/attachment?id=EdmoFBoWVz&name=supplementary_material)
Submission Number: 10067

Official Review of Submission10067 by Reviewer KqZv

Official Review · Reviewer KqZv · 22 Jan 2024, 04:05 (modified: 23 Jan 2024, 14:13)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer KqZv, Authors

Revisions (/revisions?id=H6Hh8HtQAi)

Paper Summary:
For the change detection task, this paper focuses on the problems of previous state-of-the-art methods, high false positive rate, and lack of explainability. It proposes to use a 1-head, co-attention layer at the end of a Transformer encoder to serve as an explainable and editable attention bottleneck. This paper conducts abundant experiments and visualization to show the advancement and explainability of the proposed method.

Paper Strengths:
- The paper is well organized. The main idea of the paper is easy to catch. Some figures are clear.
- The experiment results show the effectiveness of the proposed method among state-of-the-art methods and the visualization shows examples of reducing false positive rate.

Paper Weaknesses:
- I agree with the motivation of this paper to a certain extent, but the relationship between the motivation and the proposed method is weak. The author should provide more analysis about the focused questions, for example, analyzing the cause of the high false positive rate and explaining the justification of the proposed method for explainability and editable.
- The author uses a transformer encoder with self-attention layers to fuse the image features from two images ahead of the proposed co-attention bottleneck. The input tokens of the bottleneck are fused results of two images, which could not be considered as independent features from any one image. Even though the author uses f1 and f2 features as the query input and k/v input of the co-attention bottleneck, it is hard to understand and explain the real role they play.
- The author conducts the ablation experiments of ZeroCLS and ZeroAttention, which seem to be important in introducing the contribution of the proposed method. However, it is a little difficult to understand what's the purpose of this experiment due to the unclear description of implementation.

Overall Recommendation: 2: weak reject
Justification For Recommendation And Suggestions For Rebuttal:
- I suggest authors support more information about the analysis for focused problems, the modeling of the proposed method, and the relationship between them.
- In my opinion, this paper only proposes a simple modification by changing the multiple-head self-attention layer to a 1-head co-attention layer. From the view of technology, I think the contribution to the community is limited. The author should provide more convincing analysises, experiments, or comparisons to demonstrate the creativity or necessity for the proposed method.

Confidence Level: 4: The reviewer is confident but not absolutely certain that the evaluation is correct.

Official Review of Submission10067 by Reviewer AVM1

Official Review · Reviewer AVM1 · 22 Jan 2024, 01:23 (modified: 23 Jan 2024, 14:13)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer AVM1, Authors

Revisions (/revisions?id=XVswdy3XbL)

Paper Summary:
This paper proposes a framework to detect image changes by leveraging Transformer encoders and co-attention bottleneck. The proposed approach perform well on the selected benchmark, especially for the image pairs without change.

Paper Strengths:
Overall, the paper is well written and the proposed method is straightforward. The figures and tables in the papers greatly assist the readers to understand the paper.

Paper Weaknesses:
The proposed method has better performance than other mentioned approaches on no-change image pairs. However, for the image-pairs with changes, the proposed method may not have too much advantages over its counterparts. Do you analyze the reason for that?

The paper stresses on explainability of the method. but I still do not understand how and why the proposed method is more explainable than other methods. It could be better if the authors make it clearer why the proposed method is explainable and other method is not that explainable.

To compare to CYWS, the backbone is ViT v.s. U-Net. Do you consider using consistent backbone to compare the methods?

Overall Recommendation: 3: borderline
Justification For Recommendation And Suggestions For Rebuttal:
Please refer to paper weakness.

Confidence Level: 4: The reviewer is confident but not absolutely certain that the evaluation is correct.

Official Review of Submission10067 by Reviewer ABUN

Official Review Reviewer ABUN 13 Jan 2024, 10:47 (modified: 23 Jan 2024, 14:13)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer ABUN, Authors

Revisions (/revisions?id=xDtOGH78Wm)

Paper Summary:

1. Issue in Existing Model:
    - The abstract addresses the challenge in existing Vision Transformer (ViT) models where attention maps are not self-explanatory. They require a post-hoc feature attribution method for aggregation into a single heatmap, making the interpretability of attention maps complex.
2. Proposed Solution:
    - The proposed solution introduces a 1-head Co-Attention layer placed at the end of a Transformer encoder, just before the final classification head. This layer acts as an exact, explainable, and editable attention bottleneck. The objective is to enhance the interpretability of attention maps in ViTs.
3. Evaluation on Various Standard Datasets:
    - The authors evaluate the effectiveness of their proposed solution by training the model to predict object-level changes between two images, a task referred to as image difference prediction. The co-attention bottleneck of their binary classifier exhibits superior performance in spotting differences compared to existing weakly-supervised ViT models on this task. Additionally, it shows comparable performance to change detectors explicitly trained with bounding-box supervision.
    - The proposed co-attention bottleneck is not only effective but also editable, allowing modifications. Moreover, it demonstrates a strong causal relationship with the classifier's prediction, a unique property highlighted for the first time in the literature.

Paper Strengths:

1. Addressing False Positives:

The proposed Co-attention Bottleneck (CAB) effectively addresses the issue of high false positives in no-change

scenarios, a limitation observed in the CYWS model. It provides more accurate localization, reducing false positives and enhancing reliability.

2. Explainability and Editability:

CAB introduces a Co-attention Bottleneck layer designed for explainability and editability. Unlike existing models that rely on post-hoc feature attribution methods, CAB is self-explanatory and editable by users, allowing for more transparent and user-friendly interpretation

Paper Weaknesses:

1. Computational Complexity:
   - The paper introduces three design choices for attention bottleneck modules, but it's important to thoroughly evaluate and discuss the computational complexity of each choice. Particularly, consider the impact on training and inference times, especially with large datasets.
2. Performance Comparison:
   - While the paper briefly mentions that co-attention with exactly 1 head performs comparably to more complex bottlenecks, a detailed performance comparison table or analysis should be provided. Highlight the trade-offs in terms of performance and computational requirements.
3. Explanatory Clarity:
   - The description of the attention bottleneck modules (all-attention, masked-attention, co-attention) is detailed, but ensure that the paper maintains clarity in explaining the concepts. Provide visual aids or examples to enhance understanding.
4. Scalability and Generalization:
   - Investigate the scalability and generalization capabilities of the proposed framework. How well does it perform with larger or more diverse datasets beyond the ones mentioned in the paper? Address the potential limitations in adapting to various domains.

5.Limitations in Masked-Attention:

- Discuss the limitations or challenges associated with the masked-attention module. If it does not significantly reduce the number of multiplications compared to all-attention, provide insights into why and potential improvements.

6. Detailed Comparison with Existing Approaches:
   - Compare the proposed co-attention bottleneck with existing attention mechanisms in terms of performance, computational efficiency, and model interpretability. Include a detailed literature review to highlight the novelty and significance of the proposed approach.
7. Experimental Setup and Hyperparameters:
   - Provide a clear description of the experimental setup, including hyperparameters used for training and evaluation. Transparently report any challenges faced during the experimental process.
8. Visualization of Attention Maps:
   - If possible, include visualizations of attention maps[1] generated by the different attention bottleneck modules. This will aid in understanding how well the model captures relevant features and changes in images.
9. Scalability to Different Network Architectures:
   - Explore how well the proposed attention bottleneck modules can be integrated into different network architectures. Assess their compatibility and scalability with various vision encoders beyond the one used in the paper.
10. Potential Overfitting:
    - Discuss strategies employed to prevent overfitting, especially with complex attention modules. Consider presenting results on validation datasets and discussing the model's generalization capabilities.
11. Comparison with Baseline Models:
    - Include a comparison with baseline models or traditional methods for change detection to provide a comprehensive evaluation of the proposed framework's effectiveness.
12. Readability and Reproducibility: However, I feel that the paper misses one of the core aspects of machine learning practice: readability and reproducibility of results. What core mechanism of the proposed explanation method is not clear here. The author should provide an algorithm or pseudocode to reproduce the results, which this paper misses

Ensure that the paper addresses these points to enhance its clarity, completeness, and contribution to the field of change detection in images.

Reference: [1] Patro, Badri Narayana, Vinay P. Namboodiri, and Vijay Srinivas Agneeswaran. "Explaining the Unexplainable: A Critical Analysis of Explanation Approaches for Vision Transformers." Available at SSRN 4683758.

Overall Recommendation: 3: borderline
Justification For Recommendation And Suggestions For Rebuttal:
Please refer weakness section.

Confidence Level: 4: The reviewer is confident but not absolutely certain that the evaluation is correct.

Official Review of Submission10067 by Reviewer yP6B

Official Review Reviewer yP6B 09 Jan 2024, 15:24 (modified: 23 Jan 2024, 14:13)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer yP6B, Authors
Revisions (/revisions?id=06KrLZWuSu)

Paper Summary:
This paper aims to solve the change detection task. The authors claim high false positive and lack of explainability issues of previous change detection methods. To tackle this, the authors design a Co-attention bottleneck between a change detection layer and a backbone adopted from CLIP4IDC. While there are three applicable candidates for the bottleneck, the Co-attention bottleneck is employed due to its efficient computation.

Paper Strengths:
- The performance shows superior results in some evaluation tasks.

Paper Weaknesses:
- Insufficient contributions
  - The claimed goal is not persuasive enough. The authors should analyze explicit negative impacts of excessive false positive predictions.
  - Framework design
    - Co-attention is almost identical to self-attention except values and keys are derived from second view (f^2)
    - The modules in Fig. 2(b) are borrowed from CLIP4IDC and is not a contribution of the paper.
- Insufficient description and explanation
  - Is there an explicit explanation on how the proposed framework design contributes to the reduction of false positives? Figure 2 only describes classification process for change detection.
  - The framework in Fig. 2 seems not to describe bounding box prediction task.
- Experiments
  - Comparison between the proposed method and CYWS seems not fair. The proposed modules adopt some layers from CLIP, pre-trained by enormous external datasets. Moreover, the backbone are also different (UperNet v.s. ViT-B).
  - In lines 207-208, the authors explained the strength of Co-attention as the reduced multiplication. However, there is no comparison on computational costs among the introduced three attention blocks.
- Writing issue
  - Paper is not written well overall
  - Subsection titles in Section 4 are too wordy. They can be more concise.
  - It would be more appropriate to move Section 3.4 to Section 4
  - Erratum
    - line 51, no space between . and OpenImages-C
    - The term "transformer encoder blocks" is inappropriate. I recommend to use "transformer encoders"

Overall Recommendation: 2: weak reject
Justification For Recommendation And Suggestions For Rebuttal:
The paper should be revised and re-organized overall. I'll rate this paper by weak reject. Please refer to the weakness part above.

Confidence Level: 3: The reviewer is fairly confident that the evaluation is correct.