MMC Transformer: Multiscale Memory Comparator Transformer for Few-Shot Video Segmentation

Anonymous

05 Nov 2022 (modified: 03 Mar 2023) Submitted to CVPR 2023 Readers: Conference, Paper5868 Senior Area Chairs, Paper5868 Area Chairs, Paper5868 Reviewers, Paper5868 Authors Show Bibtex Show Revisions

Abstract: Learning to compare support and query features in few-shot video transformers has been shown to be a powerful approach. Typical approaches limit comparisons to a single feature layer and thereby ignore potentially valuable information: Comparators that operate with early network layer features support precise localization, but lack sufficient semantic abstraction; at the other extreme, operating with deeper layer features provides richer descriptors, but sacrifices localization. We address this scale selection challenge with a meta-learned Multiscale Memory Comparator (MMC) video transformer that combines information across scales. We present novel multiscale memory transformer decoding within a meta-learning framework. This augmented memory preserves the detailed feature maps during information exchange across scales and reduces confusion between the background and novel class. Integral to the approach, we investigate multiple forms of information exchange across scales in different tasks and provide insights with empirical evidence on which to use in each task. The overall comparisons among query and support features benefit from both rich semantics and precise localization. We demonstrate our approach primarily on few-shot video object segmentation and an adapted version on the fully supervised counterpart. To further show our generality, we also extend the approach to actor/action segmentation. In all cases, our approach outperforms the baseline and/or yields state-of-the-art performance.

Supplementary Material: zip

Revealed to Mennatullah Siam, Rezaul Karim, He Zhao, Richard Wildes

31 Oct 2022 (modified: 03 Mar 2023) Submitted to CVPR 2023

Authors: *Mennatullah Siam, Rezaul Karim, He Zhao, Richard Wildes*

Authors Confirmed: I understand and agree to the following: After the registration deadline (Nov. 4), 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. 11. 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/2023/OpenReviewAuthorInstructions. Papers with one or more authors without an updated OpenReview profile by Nov. 11 may be desk-rejected.

Student Paper: No

Closest Subject Area That Your Submission Falls Into: Transfer, meta, low-shot, continual, or long-tail learning

Guidelines Confirmed: I confirm that I checked and agree to the author (https://cvpr.thecvf.com/Conferences/2023/AuthorGuidelines) and ethics guidelines (https://cvpr.thecvf.com/Conferences/2023/EthicsGuidelines).

Reply Type: all Author: everybody Visible To: all readers Hidden From: nobody **6 Replies**

Paper Decision

CVPR 2023 Conference Program Chairs

27 Feb 2023 CVPR 2023 Conference Paper5868 Decision Readers: Program Chairs, Paper5868 Senior Area Chairs, Paper5868 Area Chairs, Paper5868 Reviewers Submitted, Paper5868 Authors

Decision: Reject

Comment: In this paper, the authors propose a method called MMC Transformer for few-shot video segmentation. All three reviewers find the writing confusing or hard to follow. The reviewers also raise concerns regarding novelty and comparison to prior works. The authors provide a rebuttal in response to the questions, but the reviewers' concerns are still not fully addressed. After discussions among the AC-triplet, ACs agree with the concerns. The authors are encouraged to consider the feedback provided by the reviewers and revise the paper accordingly to improve the clarity and impact of this paper.

Comment on R3

CVPR 2023 Conference Paper5868 Authors *Mennatullah Siam (privately revealed to you)*

30 Jan 2023 CVPR 2023 Conference Paper5868 Confidential Comment to AC Readers: Program Chairs, Paper5868 Senior Area Chairs, Paper5868 Area Chairs, Paper5868 Authors

Comment:

Dear AC, we kindly point out that reviewer three has not carefully read the paper and has missed important details from the paper by not reviewing it properly. This is validated in two aspects: (i) requesting run-time and training memory although we do provide it in the supplementary and refer to its existence in the main submission L(405), L(639). (ii) providing the main justification of rejection as writing is hard to follow although the details he discusses in the weaknesses are either minor comments or are already addressed in the submission as clarified in the rebuttal. Other reviewers have agreed on identifying the main contributions behind our work as detailed in our rebuttal.

Rebuttal by Paper5868 Authors

CVPR 2023 Conference Paper5868 Authors *Mennatullah Siam (privately revealed to you)*

30 Jan 2023 CVPR 2023 Conference Paper5868 Rebuttal Readers: Program Chairs, Paper5868 Senior Area Chairs, Paper5868 Area Chairs, Paper5868 Reviewers Submitted, Paper5868 Authors

Official Review of Paper5868 by Reviewer ghUr

CVPR 2023 Conference Paper5868 Reviewer ghUr

15 Jan 2023 (modified: 07 Feb 2023) CVPR 2023 Conference Paper5868 Official Review Readers: Program Chairs, Paper5868 Senior Area Chairs, Paper5868 Area Chairs, Paper5868 Reviewers Submitted, Paper5868 Authors

Paper Summary:

The paper presents, Multiscale Memory Comparator Video Transfromer (MMC video transformer), a meta-learning based method for combining features across different scales of the backbone. MMC Transformer utilizes cross-attention between backbone features on target query and meta-learned memory features to process with a boundary refinement module. Results are reported in few-shot setting in Youtube VIS and also on full supervised action/actor segmentation where they seem to achieve significant margins over prior counterparts.

Paper Strengths:

+ **Strong Results across several settings**: The authors report strong result gains from their method across several benchmarks (youtube VIS, AVOS , DAVIS MoCA etc.) and on several common metrics that rigorously show the benefits of using MMC video transformer. The setting appear to be fair but I am not very familiar with the evaluation procedure in few-shot video segmentation.

+ **Well-ablated design decisions**: Design decision to use multi-grid, learning of memory K and V, also the number of queries are well ablated that allow the reader to gain insight on the design decisions. Additional ablations presented in supplementary on input resolution and frequencies is also welcome.

+ **Qualitative Results** : Video qualitative results demonstrate clear improvement on the multi-scale baseline.

Paper Weaknesses:

-- **Paper Framing and Novelty**: The paper mentions on meta-learning the memory module while in practice, the memory module are simply key/query vectors that are learnt during standard training with back-propagation. No particular details of a meta-learning algorithm seem to be employed. Memory seems to be referring simply to the learnt vectors that does not change according to the target query. In such a framing, all weights of a network would be considered "memory". The authors should simplify the method rather than obfuscate with non-precise framing. Adjoining to the framing, the idea of multi-grid methods for fusing localization features is quite old. While I don't believe novelty to be a big concern given the results, I do believe that authors do not do a fair job at literature review of this seminal idea.

-- **Strong appearance bias without object consistency cues** : The qualitative results supplies indicate that MMC Video Transformer suffers from strong appearance bias, where similarity in appearance seems to dominate results disregarding motion or object position consistency cues (for example in the "adult jumping" video result in supp video).

--- **Potentially weak baseline setup**: In my understanding, the stacked scale feature mixing approach is rarely employed, and the multi-grid ideas are quite commonplace in standard object detection/segmentation community. In such a case, comparison with stacked baseline as a guiding north star seems to be setting up a strawman baseline.

Minor suggestions:

- L634 - L636:
  > On Fold 3 our boundary accuracy is worse than DANet, due to challenging scenarios where the objects undergo occlusion or motion blur. this can be a satisfactory explanations for absolute results but certainty the baseline faces a similar challenge with the difficult data. Hence, this is a non-explanations for the observed trend.
- The bottom half of Fig 3 is quite difficult to understand, both on its own and even after reading the supporting text.
- The result section makes a lot of conclusions on design decision and baselines but is not organized very well. Grouping by data source / hypothesis to be validated is encouraged.

Overall Recommendation: 3: borderline

Justification For Recommendation And Suggestions For Rebuttal:

While there are issues with paper framing and some baseline setups, the reported results seem significant and the idea itself has been supported with extensive testing and ablations. I recommend a borderline for now, learning towards a weak accept, pending authors' remarks on paper framing, result presentations and baseline choice.

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

Final Rating: 2: weak reject

Final Rating Justification:

The rebuttal seems to not address my concerns, especially regarding framing the work as meta-learning. The authors seem to mean it in a way that is not used. This will cause readers further confusion in a work that is written in a way that is hard to understand as in. Further, the baseline setup seem intentionally weaker using stacked baselines and as reviewer bUmN points out, it is indeed hard to read/understand. I do not think in the current state, this is a worthwhile contribution to the community.

Official Review of Paper5868 by Reviewer aQRE

CVPR 2023 Conference Paper5868 Reviewer aQRE

03 Jan 2023 (modified: 08 Feb 2023) CVPR 2023 Conference Paper5868 Official Review Readers: Program Chairs, Paper5868 Senior Area Chairs, Paper5868 Area Chairs, Paper5868 Reviewers Submitted, Paper5868 Authors

Paper Summary:

This paper presents a method called the Multiscale Memory Comparator (MMC) transformer for few-shot video dense prediction tasks, such as video object segmentation and actor/action segmentation. The MMC transformer is a meta-learned model that combines information from multiple scales, or layers, in a transformer network in order to improve performance on these tasks. It does this by using a multiscale memory transformer decoding scheme that preserves the detailed feature maps and reduces confusion between the novel class (the class being segmented in a particular example) and the background. The MMC transformer is shown to outperform the baseline and achieve state-of-the-art performance on the tasks tested.

Paper Strengths:

1). It combines information from multiple scales in the transformer network, which allows it to take advantage of both the detailed localization of shallow layers and the rich semantics of deeper layers. Figure 2 gives an detailed illustration and Table 1 and Table 2 validates it.

2). It uses a multiscale memory transformer decoding scheme that preserves the detailed feature maps and reduces confusion between the novel class and the background. Table 3 shows its gains over multi-scale baselines.

3). It investigates multiple forms of information exchange across scales and provides insights on which to use in each task.

4). It outperforms the baseline and achieves state-of-the-art performance on the tasks tested (few-shot video object segmentation, actor/action segmentation).

Paper Weaknesses:

1). The method may be more computationally expensive than some alternatives due to the use of multiple scales and a transformer architecture. Could author[s] additionally report params/flops and inference time compared to the baseline?

2). The overall model looks relatively complicated. It is not clear how sensitive the method is to hyperparameter choices or the specific architecture used. Author[s] are encouraged to report the hyperparameters used for each task in Appendix later.

3). I am not familiar with the datasets used in this paper, so I cannot give any judgement from the quantative perspective. From the provided videos in appendix, it is not clear to me how well the method would generalize to real-world situations with more complex backgrounds or variations in lighting, camera angle, etc. Maybe author[s] can provide some additional insights.

Overall Recommendation: 4: weak accept

Justification For Recommendation And Suggestions For Rebuttal:

Suggestions are listed in the weakness section.

Confidence Level: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper.

Final Rating: 3: borderline

Final Rating Justification:

I still have confusions after reading author's response, I tend to agree with other reviewers and decrease my score accordingly.

Official Review of Paper5868 by Reviewer bUmN

CVPR 2023 Conference Paper5868 Reviewer bUmN

28 Dec 2022 (modified: 07 Feb 2023) CVPR 2023 Conference Paper5868 Official Review Readers: Program Chairs, Paper5868 Senior Area Chairs, Paper5868 Area Chairs, Paper5868 Reviewers Submitted, Paper5868 Authors

Paper Summary:

This paper works on the problem of few shot video segmentation. There are two contributions as reflected in the title, multiscale information exchange in a transformer and memory decoding in a meta-learning framework. Results on several datasets are solid and showcase the effectiveness of the proposed approach

Paper Strengths:

1. Experimental results seem strong.
2. The proposed technique is reasonable and solid.

Paper Weaknesses:

1. Writing can be improved. I listed several feedback below, but there are more I don't follow.

(1) There are many components in the system, Figure 2 alone is not clear to readers, at least not clear to me. For D_{hypercorr}, D_{multiscale} and D_{refine}, they should have high-level figures as well.

(2) Abstract goes into detail too soon without introducing what is the problem and what is the challenge.

(3) I don't understand "... which is the first multiscale video transformer to compare few labelled examples and a target video." in caption of figure 1. After finish reading the whole paper, I start to get the idea of comparator.

(4) Contribution item 2 and 3 seems duplicate.

(5) Figure 3 is hard to follow, I feel it should be divided into two figures.

(6) Decoding needs to be clearly defined, whether it refers to video decoding, or decoder network, or transformer decoder with cross attention, etc. There are many things called decode, which is quite confusing.

2. In terms of novelty, both the idea of using multiscale information and memory module has been investigated in MViT, MViTv2, and MemMViT. Although using them for video segmentation is a contribution, I find it overclaimed.

Multiscale vision transformers, ICCV 21 MViTv2: Improved multiscale vision transformers for classification and detection, CVPR 22 MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition, CVPR 22

3. One question is, how is the memory module used? Do we use it in inference or only during training?
4. Do we have run time analysis, like inference speed, training memory consumption compared to other approaches?

Overall Recommendation: 2: weak reject

Justification For Recommendation And Suggestions For Rebuttal:

Writing is hard to follow, see weakness for suggestions in rebuttal.

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

Final Rating: 2: weak reject

Final Rating Justification:

Thank you for preparing the rebuttal, it helps addressing some of my concerns but not all. Overall, the paper needs significant efforts to organize and rewrite. Because at this moment, several concepts like meta-learned is not easy to understand. The contribution from multigrid feature fusion is limited in my opinion. Hence, I keep the original rating.