Differentiable Soft-Masked Attention

Ali Athar, Jonathon Luiten, Alexander Hermans, Deva Ramanan, Bastian Leibe
RWTH Aachen University (Germany), Carnegie Mellon University (USA)

Motivation

An important component of Transformer architectures is the use of **cross-attention**. Recently **masked cross-attention** was introduced in the Mask2Former architecture, limiting the attention to a specific region in the image in a binary fashion. This improves various image and video segmentation tasks.

We propose a new **differentiable soft-masked attention** variant with two advantages:

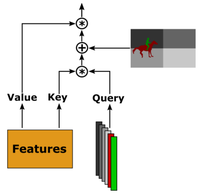
- It allows **soft, non-binary masks**
- It is **fully differentiable** with respect to the mask values

We demonstrate the effectiveness of this soft-masked attention in the context of weakly supervised **video object segmentation** (VOS)

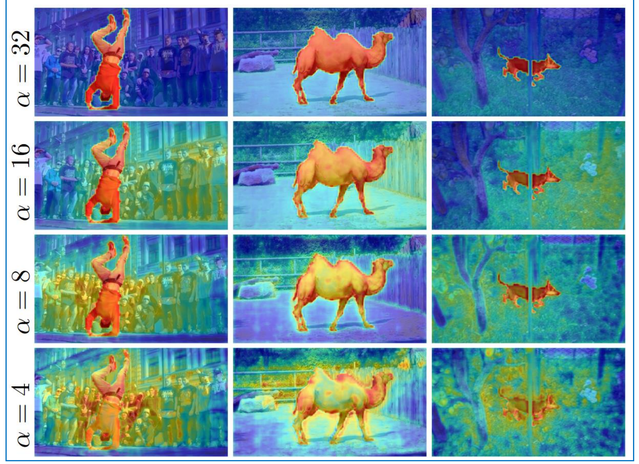
Differentiable Soft-Masked Cross-Attention

During cross-attention, every query is assigned a soft mask, which is used to condition the attention to a specific image region.

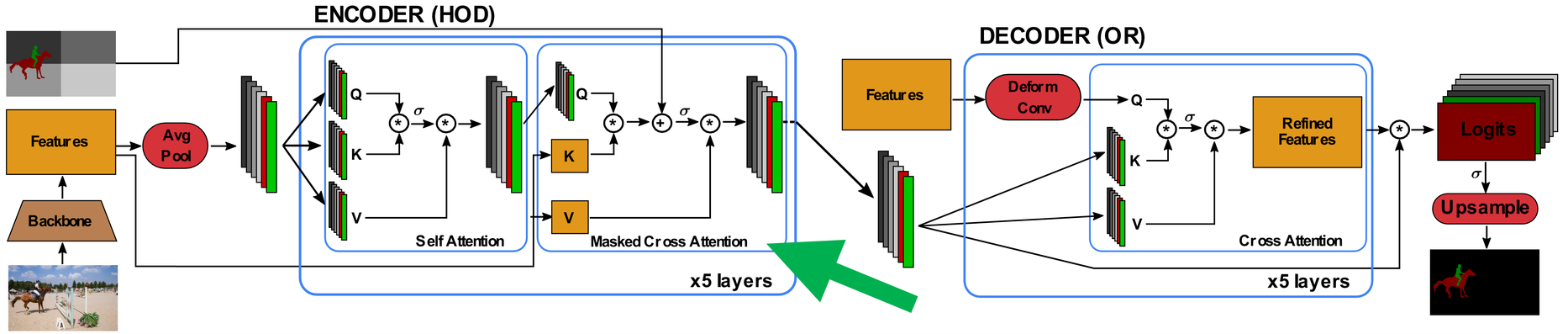
- A learnable factor α tis added o the attention matrix
- Each attention head is initialized with a different α



$$\texttt{softmax}\left(\frac{K^T Q + \alpha M}{\sqrt{C}}\right) \cdot V$$



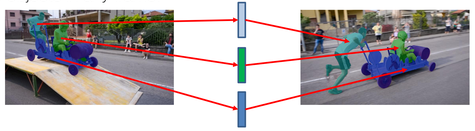
HODOR Architecture



Video Object Segmentation Approach

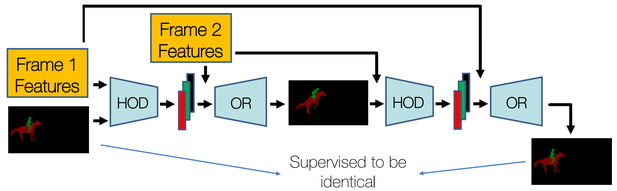
We empirically show the benefits of the soft-masked cross attention within our approach for video object segmentation, HODOR: High-level Object Descriptors for Object Re-segmentation.

- It turns object masks into object descriptors that can be used to re-segment the objects in any frame.



- **Encoder (HOD):** Feature map + object masks → Object descriptors
- **Decoder (OR):** Feature map + object descriptors → Re-segmented objects

HODOR can make use of varying training setups and unlike most other methods can be trained on image annotations only, without the need for dense video annotations. It can also utilize cyclic consistency within videos for training.



Quantitative Results

Both when training on image or video datasets, our soft masking approach outperforms baselines without masking or hard masking (meaning $-\infty$ is added to the attention matrix for areas outside of the binarized object mask).

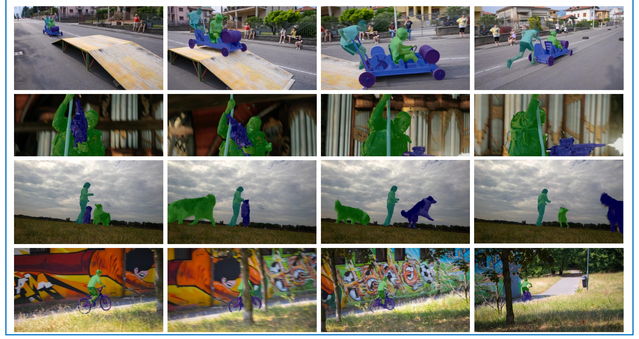
- Surprisingly, no masking is on par or even better than hard masking
- The learned soft attention masking allows the network to learn how much context to use.

Training Data	Attention Type	DAVIS val
COCO	No Masking	74.4
COCO	Hard Masking	74.5
COCO	Soft Masking (Ours)	77.5
YTVOS + DAVIS	No Masking	79.1
YTVOS + DAVIS	Hard Masking	78.4
YTVOS + DAVIS	Soft Masking (Ours)	80.6

Compared to other methods trained on image annotations only, HODOR achieves state of the art results on relevant VOS datasets.

Method	DAVIS val	DAVIS testdev	YT-VOS val
STM Oh et al. ICCV'19	60.0	-	69.1
DMN+AOA Liang et al. ICCV'21	67.9	-	-
KMN Seong et al. ECCV'20	68.9	-	-
STCN Cheng et al. NeurIPS'21	75.8	51.7	69.4
HODOR (Ours)	77.5	65.0	71.7
HODOR (Ours with cyclic consistency)	80.6	66.0	72.4

Qualitative Results (DAVIS'17)



Conclusion

We propose a differentiable soft-masked attention mechanism and demonstrate how it can be use to improve our video object segmentation approach.

We expect this to be a useful tool in many attention-based architectures, since it forces the network to focus on specific image regions, but also enables the network to learn to which extend this is meaningful.