Note: Discovering Objects that Can Move

Sinkoo

April 17, 2022

1 Introduction

Resent work on object recognition relies on expensive manual labeling, and can only cover limited amount of labels. The reason this task challenging is mostly due to the inherently ambiguous notion and context dependence. The authors proposed a method that train with videos and can segment objects in static images. This paper also utilized a new dataset for the model consists of hundreds of videos and comes with a full set of ground truth annotations.

2 Method

2.1 Background

2.2 A framework for object discovery in videos

Takes a sequence of videos frames $\{I^1, I^2, ..., I^T\}$ as input. Following is a CNN to obtain a individual frame representation $H^t = f_{enc}(I^t)$. These individual representations are aggregated by a ConvGRU spatio-temporal memory module to obtain video encoding via $H^{'t} = ConvGRU(R^{t`1}, H^t)$, where $R^{t`1} \in R^{H'*W'*D_{inp}}$ is the recurrent memory state. Then proceed to map the representation to set of slots S^t. This paper changed a little when using slot attention. The authors only perform a single attention operation to ndirectly compute the slot state $S^t = W^{t^T} v(H^{'t})$, where the attention matrix W^t is computed using the slot state in the previous frame St1. For the first frame using a learnable initial state S_0.

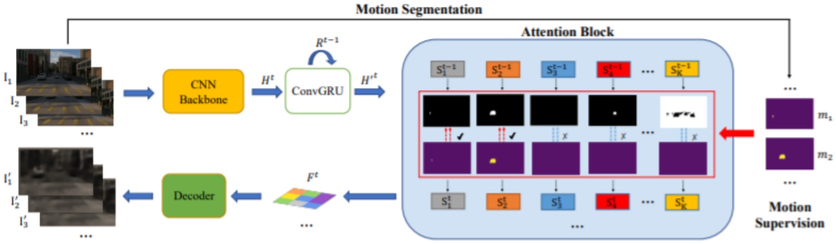

Figure 2. Our method takes a sequence of frames as input and processes them individually with a backbone network (shown in yellow), and a ConvGRU recurrent memory module. The resulting feature maps H'^t are passed to the attention module (shown in blue) which binds them to a fixed set of slot variables via an attention operation. We additionally use automatically estimated motion segmentation to guide the attention operation for a subset of the slots. Finally, the slot states are combined in a single feature map F^t and decoded to reconstruct the frame. The reconstruction objective enforces generalization from moving to static instances.

Since a full image reconstruction is expensive, this paper deal with that by broadcast slot feature to feature map F^t and use the attention mask W of the slot as an alpha mask A. Then use a decoder to reconstruct I'^t.

3 Incorporating independent motion priors