Attentive Semantic Video Generation using Captions

Abstract:

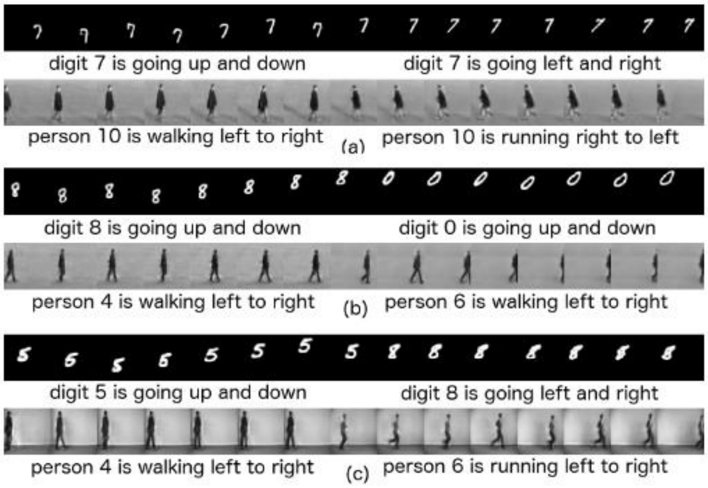
This paper proposes a network architecture to perform variable length semantic Video generation using captions.

Notes:

Since a video can be arbitrarily long, generation of such videos necessitates a step –by- step generation mechanism. Their model approaches video generation iteratively by creating one frame at a time, and conditioning the generation of the subsequent frames by the frames generatated so far. Also, every frame is itself an amalgamation of several objects moving and interacting with each other. In order to generate such frames, they follow a recurrent attentive approach, which focuses on one part of the frame at each time step for generation, and completes the frame generation over multiple time steps. By iteratively generating the frame over a number of time-steps in response to a given caption, our network adds caption- driven semantics to the generated video. A key advantage of following such an approach is the possibility to generate videos with multiple captions and thus change the contents of the video midway according to the new caption.

This paper makes the following contributions: (1) A novel methodology that can perform variable length semantic video generation with captions by separately and simultaneously learning the long-term and short-term context of the video; (2) A methodology for selectively combining information for conditioning at various levels of the architecture using appropriate attention mechanisms; and (3) A network architecture which learns a robust latent representation of videos and is able to perform competently on tasks such as unsupervised video generation and action recognition.

A major advantage of such a recurrent attention mechanism, in the context of this work, is that the network learns a single distribution that can distinguish between different elements of a frame, and attend to them in each time step. This further enables the network to dynamically combine these elements during inference to effectively generate frames.


digit 7 is going up and down
digit 7 is going left and right
person 10 is walking left to right
(a)
person 10 is running right to left
digit 8 is going up and down
digit 0 is going up and down
person 4 is walking left to right
(b)
person 6 is walking left to right
digit 5 is going up and down
digit 8 is going left and right
person 4 is walking left to right
(c)
person 6 is running left to right

patio-Temporal style transfer. First caption generates for 7 frames. Second caption continues the generation from the 8^{th} frame.

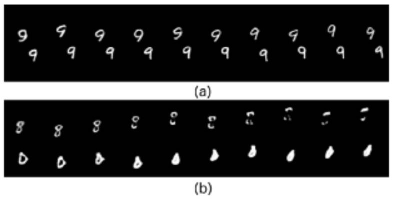


Figure 6. (a) shows videos when generated without captioning on Long-term context. (b) shows videos when generated without captioning on Short-term context.

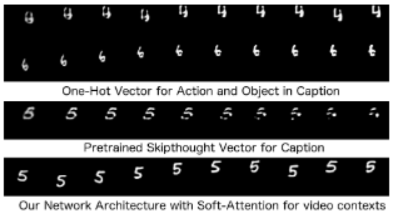

One-Hot Vector for Action and Object in Caption
Pretrained Skipthought Vector for Caption
Our Network Architecture with Soft-Attention for video contexts

Figure 7. Videos generated using different approaches of conditioning over test-set caption 'digit 5 is going up and down'