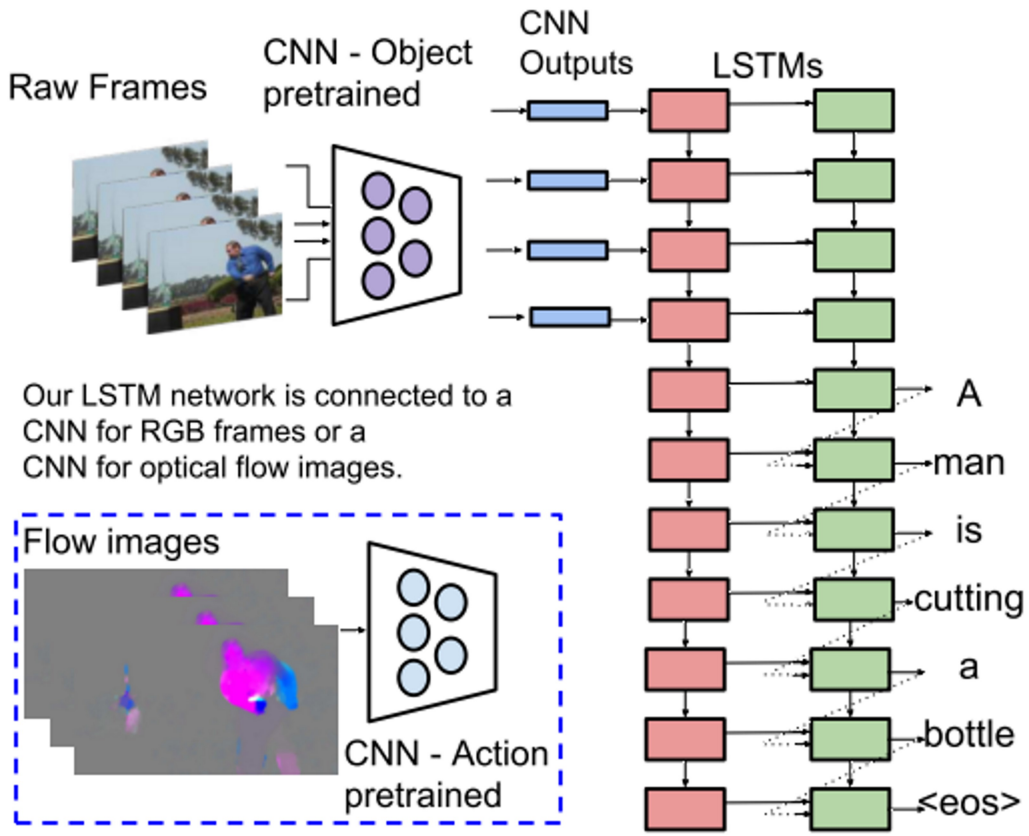
Encoder-Decoder Framework

2019年7月25日　10:52



Sequence2Sequence
CNN (VGGNet, ResNet) for spatial features
Action Recognition Models (C3D) for spatial + temporal features

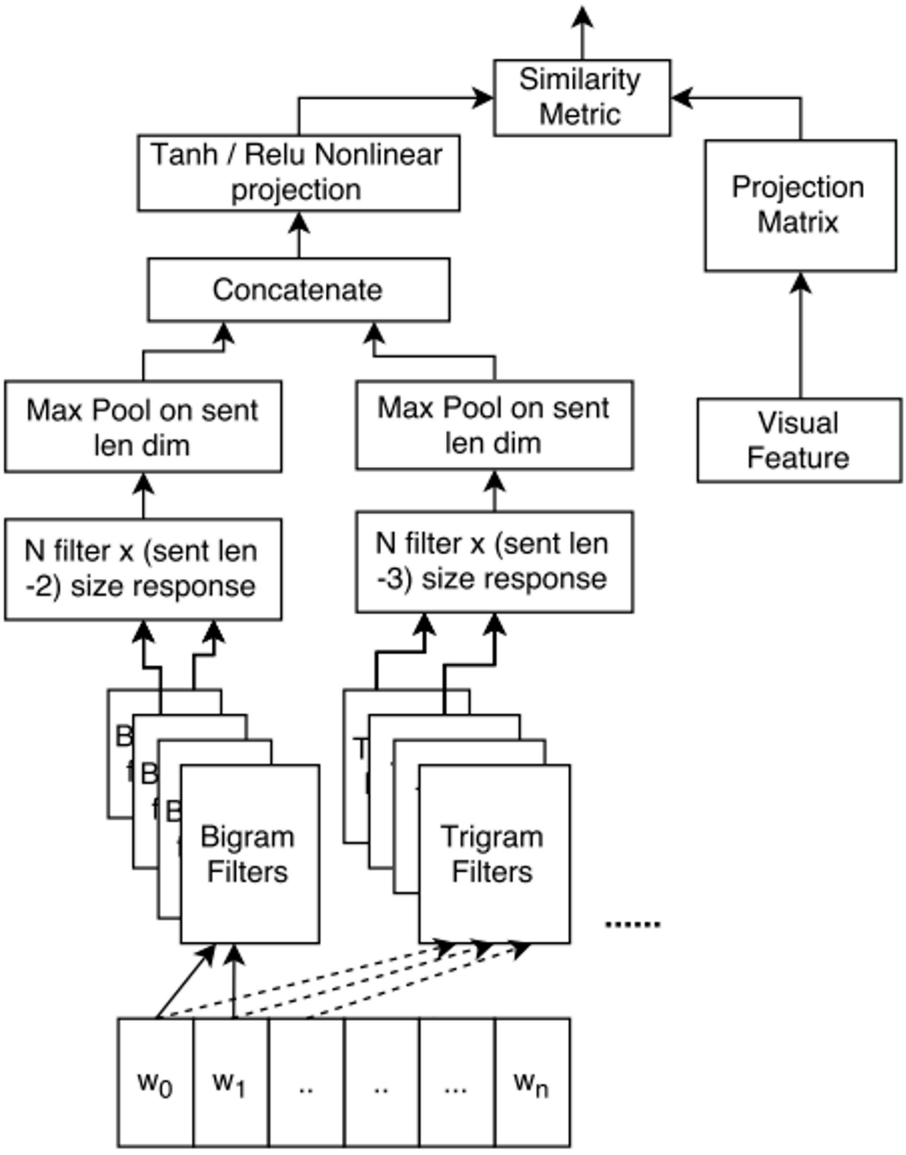
Evaluator Model

2019年7月25日 21:45

<Frame-and Segment-Level Features and Candidate Pool Evaluation for Video Caption Generation>

Motivation: each model tends to generate the best captios for different videos, through evaluator model, we can pick the best candidate and achieve better results than any single model

The model takes as input one video feature and one input sentence and computes a similarity.
It consists of a CNN sentence encoder network which takes a sequence of word vectors as input and learns a fixed-size vector embedding of the sentence.



Q: There's no necessary mapping between visual content & music sentence

Multi-Task Video Captioning

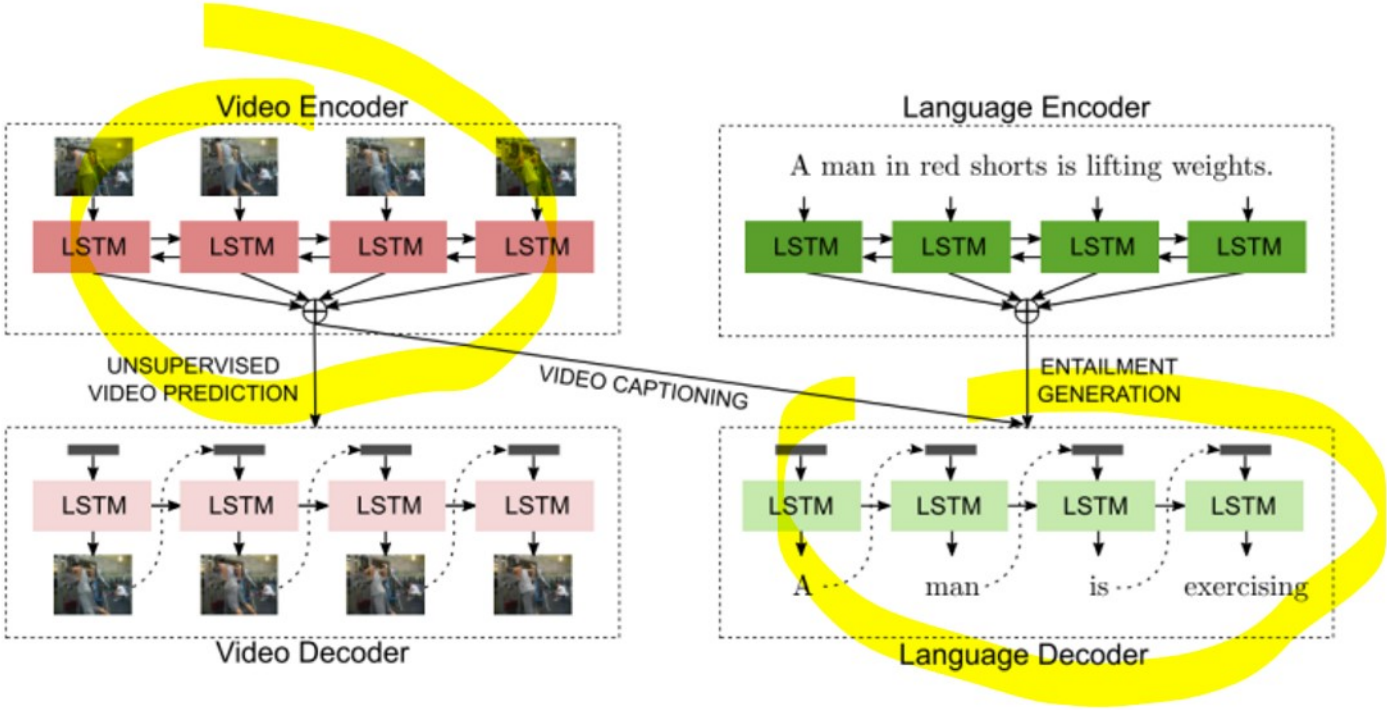
2019年7月25日 23:14

<Multi-Task Video Captioning with Video and Entailment Generation>
Accurately learning the temporal and logical dynamics involved in the task still remains a challenge, especially given the lack of sufficient annotated data.

Visual & Language representations
Temporal & logical challenge
The inadquacy of representations limits the role of encoder & decoder

Multi-task:

1. Unsupervised video prediction: using video encoder & video decoder
2. Entailment generation: using language encoder & language decoder
3. Video captioning: video encoder & languager decoder



1. Unsupervised video prediction: helps video encoder learn rich temporal representations that are aware of their action-based context and are also robust to missing frames and varying frame lengths or motions speeds
2. Entailment generation: … logical representations …
3. Multi-Task Learning: VC task shares its video encoder (parameters) with the encoder of VP task, moreover, the decoder of the video captioning task is shared with the decoder of the textual entailment generation task.

Wave
midi

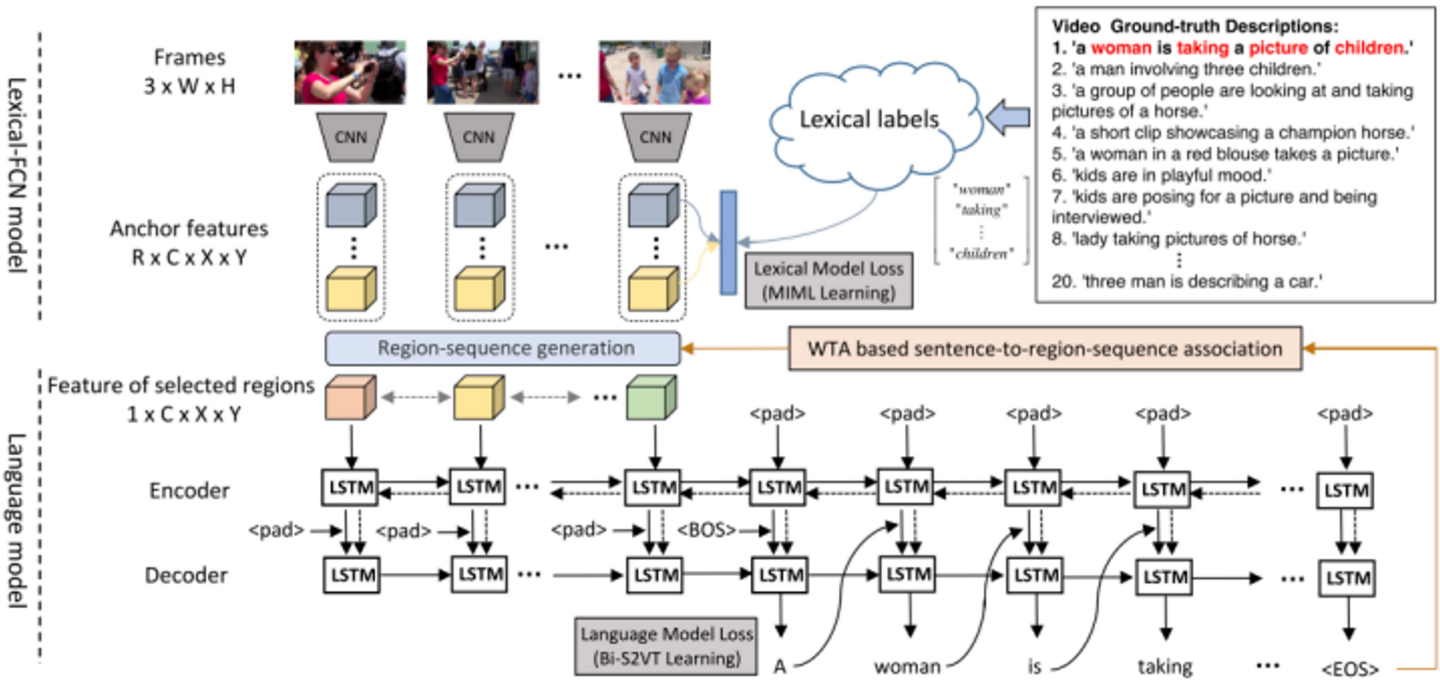
Dense Video Captioning

2019年7月25日　23:52

<Weakly Supervised Dense Video Captioning>

A single sentence cannot well describe the rich contents within images/videos.
Generating multiple sentences for different detected object localtions in videos requires region-level caption annotations for supervised training purpose.

1-to-many mapping in multiple video-level sentence annotated datasets is far from accurate.
Try to perform dense VC with associations in a weakly supervised fashion.



Multi-instance multi-label learning (MIMLL)