Co-Segmentation Aided Two-Stream Architecture for Video Captioning

Jayesh Vaidya, Arulkumar Subramaniam, Anurag Mittal

Department of of CS&E, IIT-Madras, Chennai

Problem Statement & Motivation

Problem Statement

➢ Task of generating natural language sentence by understanding local and global context.

Limitations in current works

➢ Dependence on off-the-shelf object detectors for extracting object level features.

- ○ May not exhaustively capture all object categories
- ○ Can add bias in the model.

Approach & Model Proposal

Approach

➢ Co-Segmentation based attention module (COSAM) to capture local salient regions automatically by utilizing correlation cost volume b/w adjacent frames.

➢ Salient Region Interaction Module (SRIM) consists of Object Association layer followed by self attention block to promote interactions.

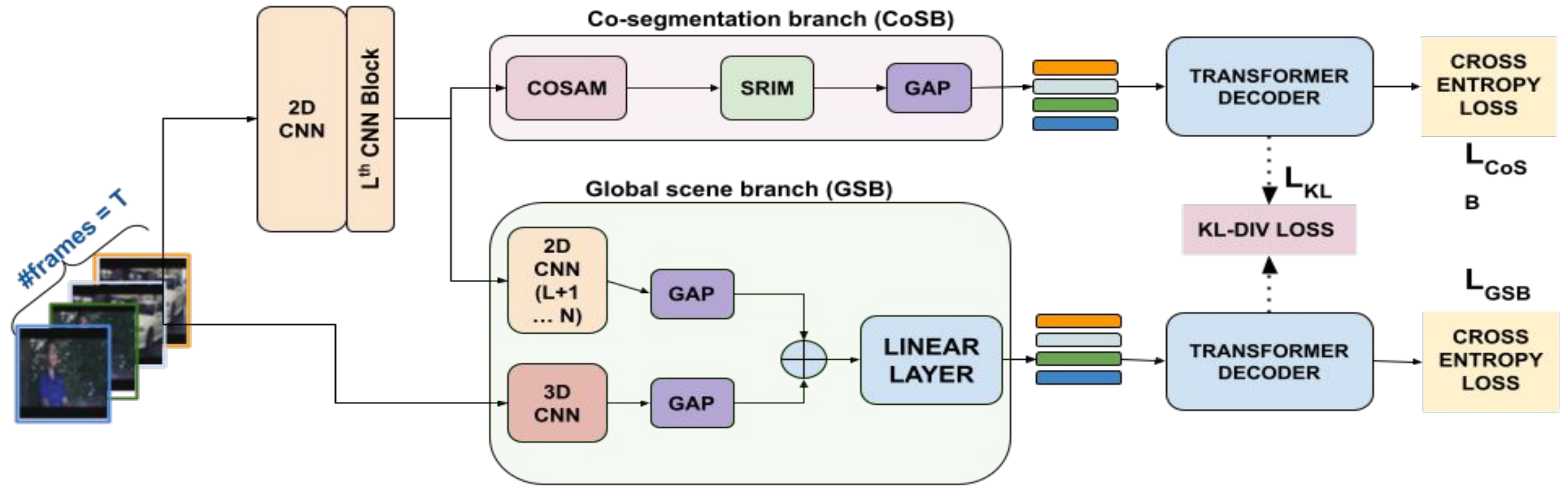
Proposal

➢ 2-branch Network, capturing local & global features, knowledge is from via KL-DIV

➢ During testing, only GSB is used.

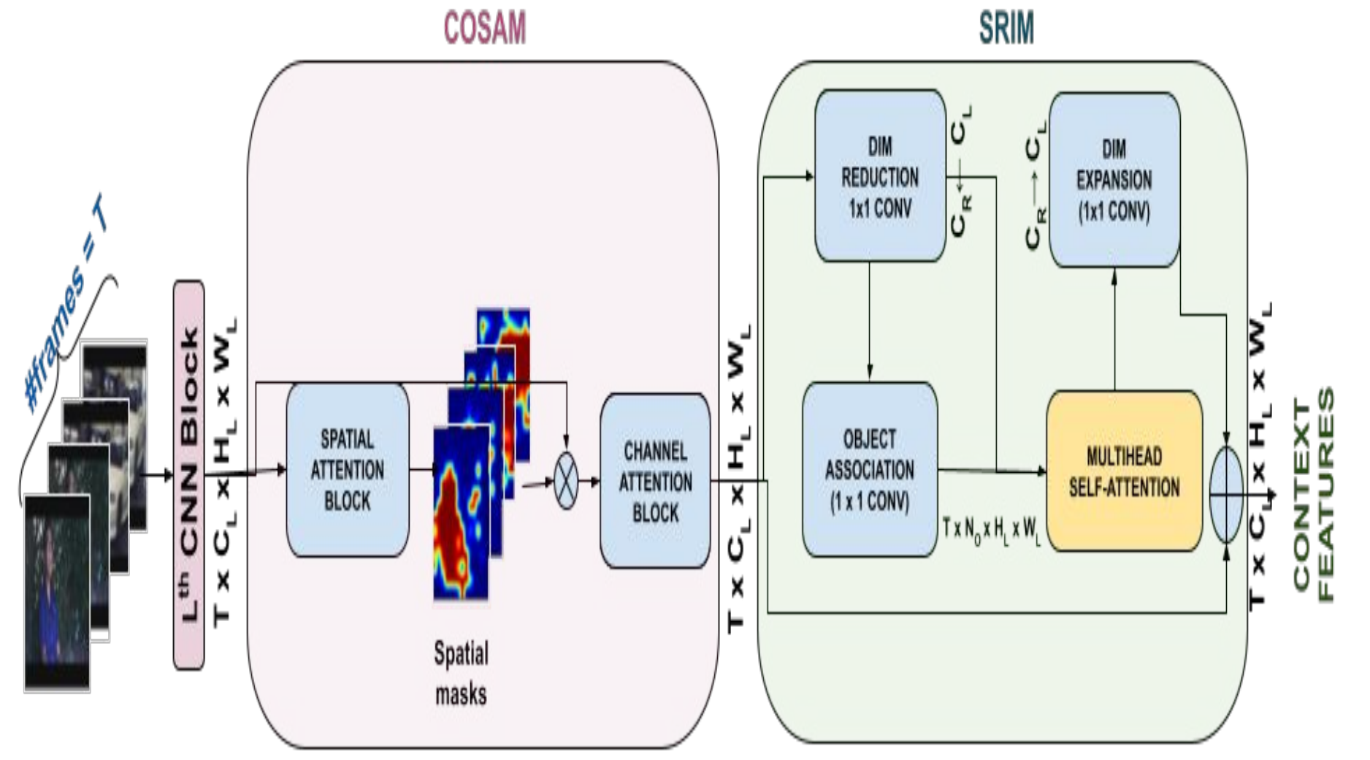
Approach & Model Proposal

Complete Architecture



➢ Intermediate feature maps are passed through COSAM attends common spatial regions.

➢ SRIM consists of multihead self attention module to promote object interactions.



Quantitative Results

Datasets - MSVD and MSRVTT

	MSVD	MSRVTT
Pan et al.	93.0	47.1
Ours	97.8	46.5

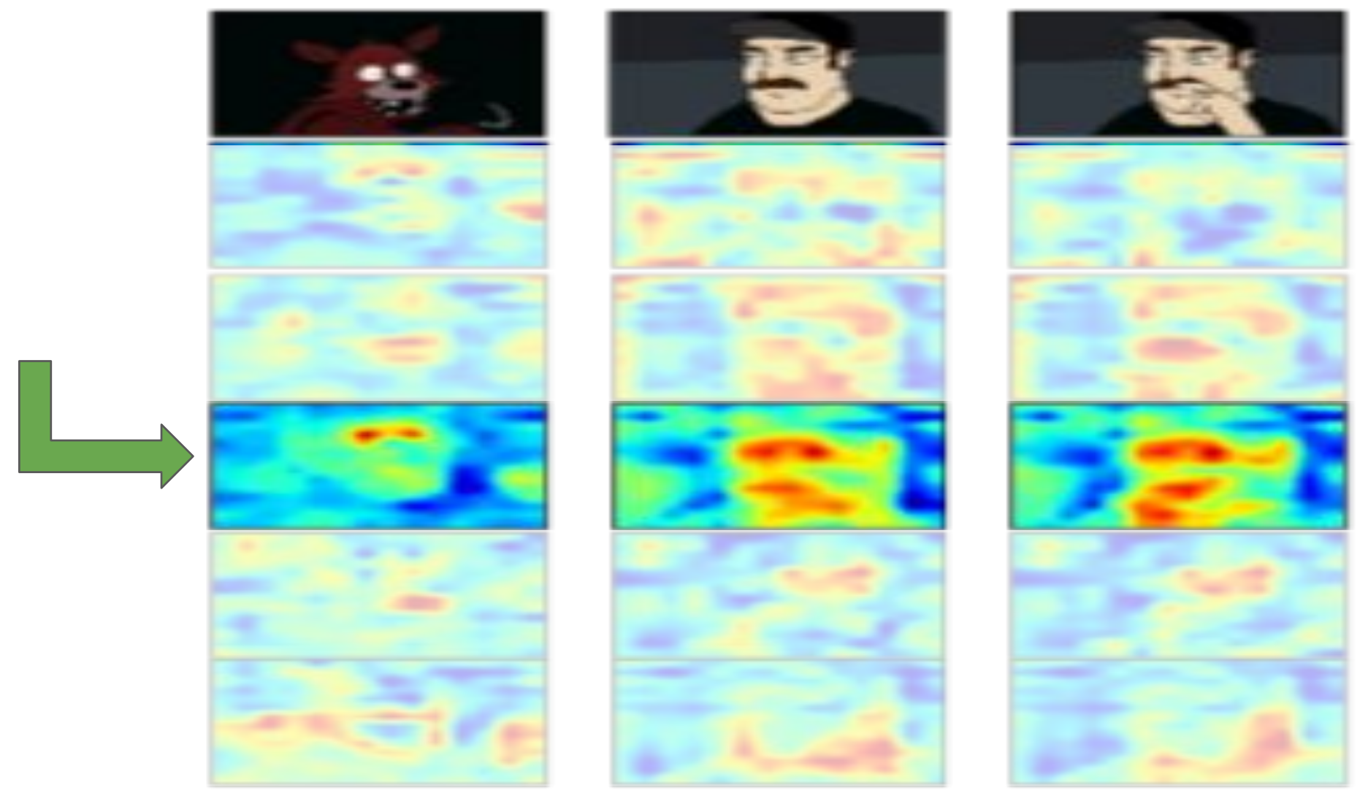
Results shown on CIDEr, we beat state-of-the-art by large margin on MSVD and get competitive results on MSRVTT.

Ablations & Qualitative Results

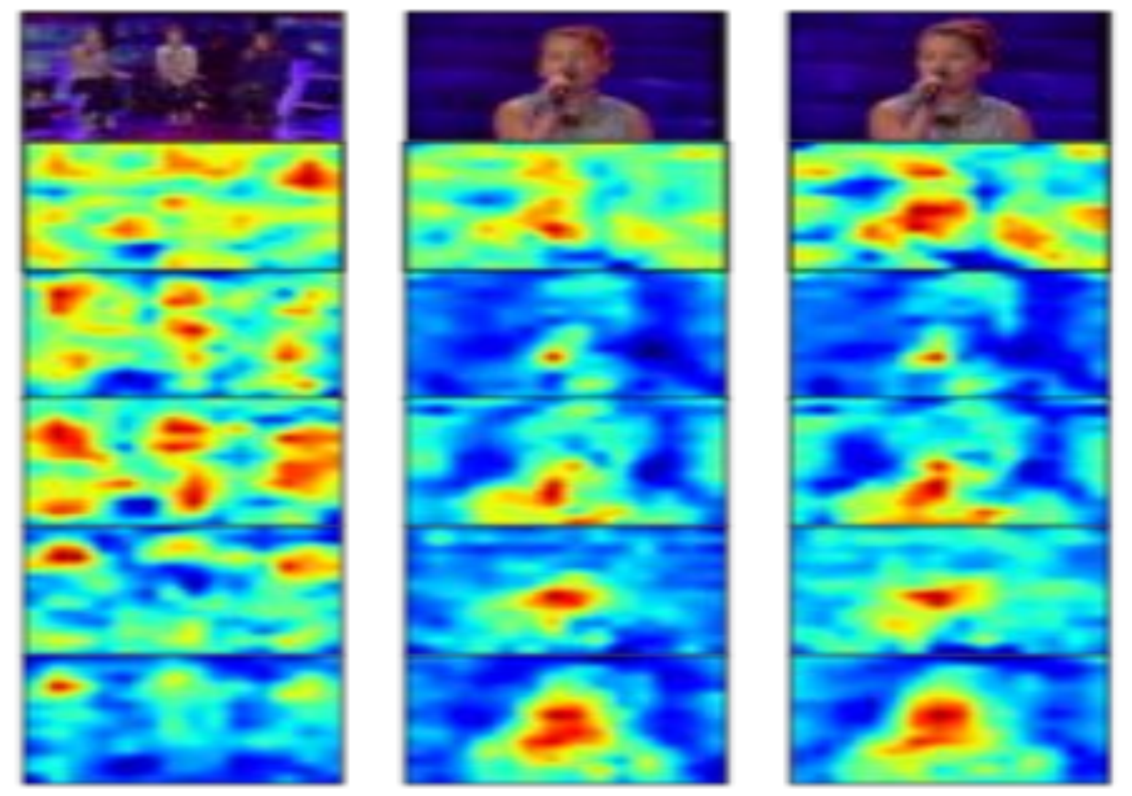
Ablations

➢ Show that COSAM+SRIM, increase the model performance significantly.

Visualizatons - SRIM and COSAM

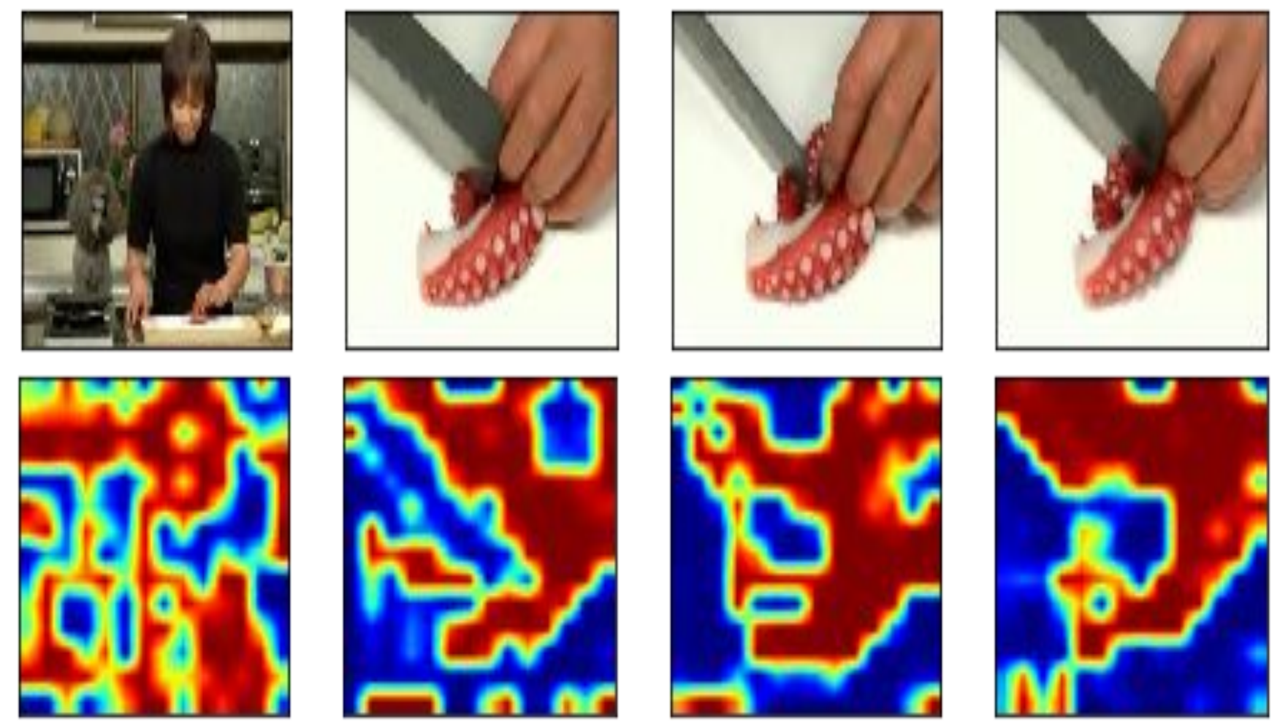


Ours: A cartoon creature is talking to a man.
GT: A cartoon with a creature is running at a man.

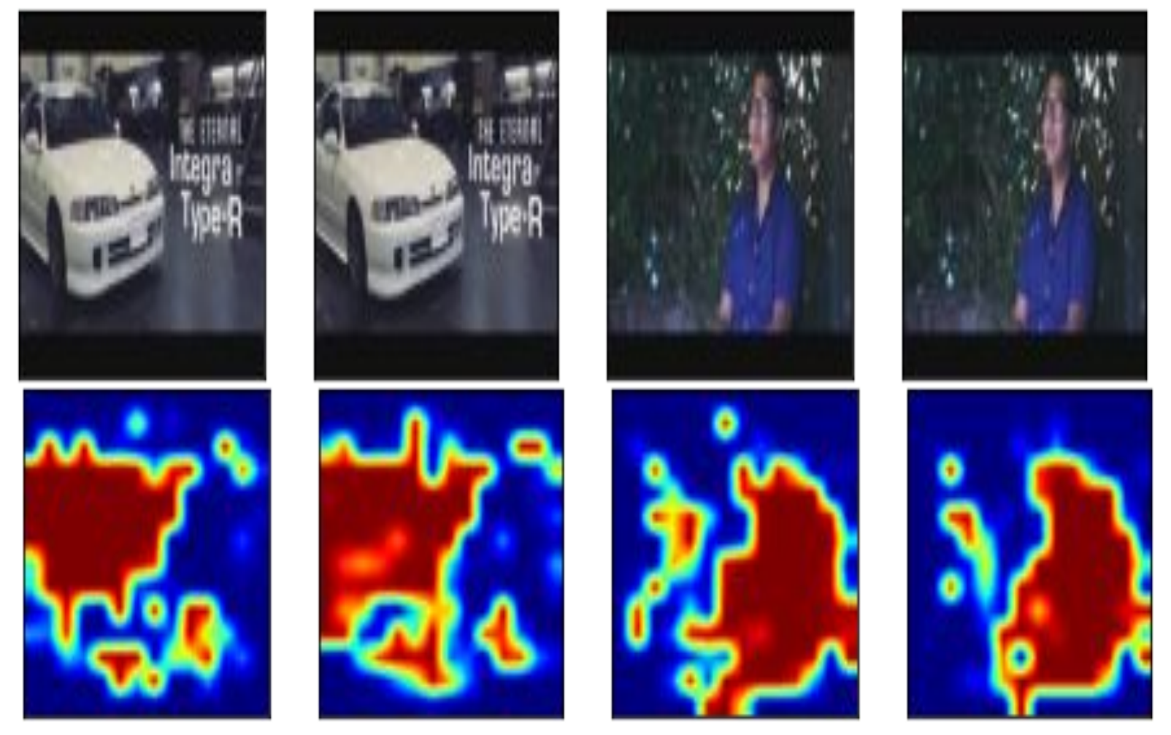


Ours: A girl is singing on a stage.
GT: A girl is singing on a stage

➢ Object detectors don't work well with the animation dataset but one the channel gets activated on eyes and mouth region of cartoon.



GT: A woman is slicing octopus.
Ours: A woman is slicing **octopus**.
STG-KD: A woman is slicing carrots.



GT: A man is talking about a car.
Ours: A man is talking about a car.
STG-KD: A man is talking about a car.

➢ Better at localizing uncommon objects.

➢ Able to detect octopus which is usually not present in object detector dataset.