Spatio-Temporal Ranked-Attention Networks for Video Captioning

[1]Anoop Cherian [2]Jue Wang [1]Chiori Hori [1]Tim Marks

[1]Mitsubishi Electric Research Labs (MERL), Cambridge, MA, [2]Australian National University, Canberra

{cherian, chori, tmarks} @merl.com jue.wang@anu.edu.au

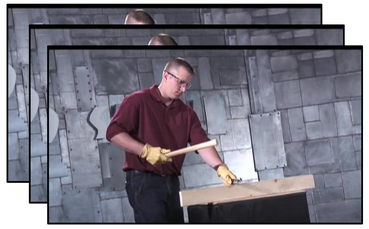
MITSUBISHI ELECTRIC RESEARCH LABORATORIES, INC

Australian National University

WACV 2020

1. Problem: Video Captioning

Given a video sequence, generate a caption (sentence) describing the video content.



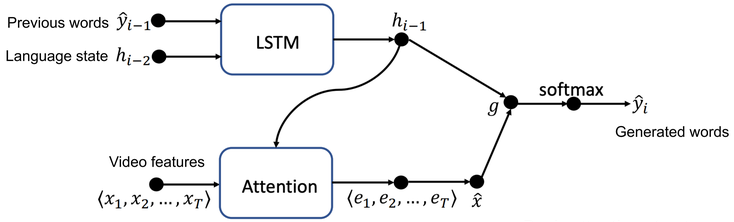
"a man is using a pipe to hammer the knife "

2. Contributions

- A novel **Spatio-Temporal and Temporo-Spatial (STaTS) Attention** scheme that attends to caption-word-specific specific cues in the input video.
  - ST influences generation of verbs/action words
  - TS helps generate nouns in the generated caption.
- A novel **Ranked-Attention scheme** that uses an LSTM to *emulate* a rank-SVM algorithm, capturing temporal order.
- An **End-to-End deep learning framework**. Experiments on two standard benchmarks demonstrating **state-of-the-art results**.

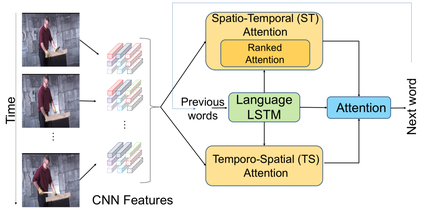
3. Prior Work

- Previous methods use spatio-temporal attention conditioned on the language state (LSTM)
  - They use the same attention mechanism for disparate types of video cues, such as static and dynamic features, objects, interactions, etc. [Zanfir et al., ACCV 17, Zhang & Peng, CVPR 19]



Basic attention structure

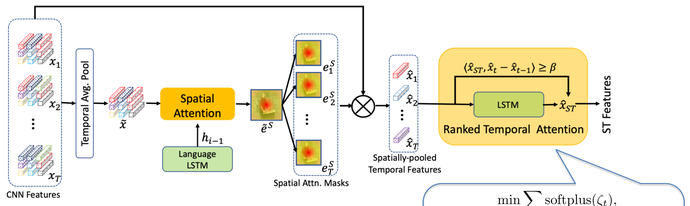
4. Spatio-Temporal & Temporo-Spatial (STaTS) Attention



Key Ideas:

- Use pre-trained 3D convolutional feature maps from every frame in the video
- Use two streams:
  a) **ST stream**, which attends to temporally varying cues (actions/verbs)
  b) **TS stream**, which chooses an individual frame, then selects regions in that frame to attend to (nouns/subject/object)
- Both streams are conditioned on the state of the language model (LSTM).

4a. Spatio-Temporal (ST) Attention



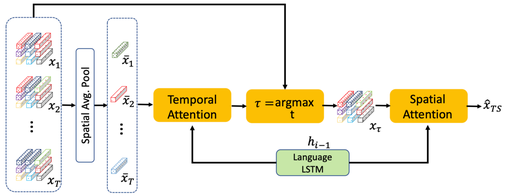
Key steps:

1. Use temporal average pooling to find spatial attention mask $\tilde{e}^S$ that applies to every frame in the video.
   - Highlights time-varying regions
2. Replicate spatial attention mask $\tilde{e}^S$ as $e_1^S, e_2^S, \ldots e_T^S$. to obtain a single feature vector $\hat{x}_t$ for each frame *t*.
3. Ranked temporal attention LSTM outputs a representation $\hat{x}_{ST}$ that respects temporal order (dynamics).
   - $\hat{x}_{ST}$ summarizes the temporal evolution of input features

$$\min_{\theta} \sum_{t} \text{softplus}(\zeta_t),$$
$$\zeta_t = \langle \hat{x}_{ST}, \hat{x}_{t-1} \rangle + \beta - \langle \hat{x}_{ST}, \hat{x}_t \rangle,$$
$$\hat{x}_{ST} = \bigoplus_{t=1}^{T} \text{LSTM}_{\theta}(\hat{x}_t).$$

Final output vector of LSTM is $\hat{x}_{ST}$
Train LSTM so that Inner product $\langle \hat{x}_{ST}, \hat{x}_t \rangle$ increases with t.

4b. Temporo-Spatial (TS) Attention



Key steps:

1. Spatially pool to generate temporal features $\bar{x}_t$
2. Conditioned on the language LSTM, using temporal attention, identify a single frame τ that will be used to generate the next caption word.
3. Apply spatial attention to this frame's features x_τ

5. Experiments and Results

MSVD Dataset: State-of-the-Art Comparisons

Scheme	CIDEr	BLEU4	ROGUE	METEOR
PickNet [11]	0.765	0.523	0.696	0.333
M^3 [56]	N/A	0.520	N/A	0.321
LSTM-LS [37]	N/A	0.511	N/A	0.326
MA-LSTM [62]	0.704	0.523	N/A	0.336
MAM-RNN [34]	0.539	0.413	0.688	0.322
RecNet [55]	0.803	0.523	0.698	0.341
GRU-EVE [2]	0.781	0.479	**0.715**	**0.350**
STaTS(FR+FL)	0.747	0.495	0.694	0.334
STaTS (I3D+FL)	**0.835**	**0.548**	0.711	**0.350**

MSR-VTT Dataset: State-of-the-Art Comparisons

Scheme	CIDEr	BLEU4	ROGUE	METEOR
Dense-Cap [46]	**0.489**	0.414	0.611	0.283
PickNet [11]	0.441	0.413	0.598	0.277
OA-BTG (R200) [73]	0.469	0.414	–	0.282
M^3-VC [56]	–	0.381	–	0.266
GRU-EVE (C3D+IVR2) [2]	0.481	0.383	0.607	**0.284**
RecNet [55]	0.427	0.391	0.593	0.266
STaTS (R152)	0.445	0.392	0.597	0.279
STaTS (R152+C3D)	0.465	0.416	**0.615**	**0.284**
STaTS (I3D)	0.434	0.401	0.604	0.275
STaTS (I3D+FL)	0.438	0.410	0.611	0.276
STaTS (I3D+FL+C)	0.451	**0.417**	0.612	0.280

5a. Ablative Studies

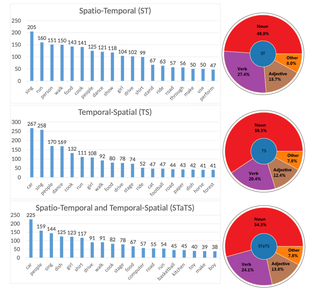
Dataset	Scheme	Feature	CIDEr	BLEU4	ROGUE	METEOR
MSVD	ST	I3D	0.742	0.502	0.68	0.325
	TS	I3D	0.521	0.391	0.646	0.289
	STaTS	I3D	**0.802**	**0.526**	**0.695**	**0.335**
	ST	FRCNN	0.686	0.477	0.69	0.33
	TS	FRCNN	0.439	0.376	0.633	0.274
	STaTS	FRCNN	**0.709**	**0.492**	**0.68**	**0.319**
MSR-VTT	ST	I3D	0.429	0.397	0.600	0.271
	TS	I3D	0.427	0.380	0.595	0.273
	STaTS	I3D	**0.434**	**0.401**	**0.604**	**0.275**

Comparing attention schemes and feature types.

Scheme	CIDEr	BLEU4	ROGUE	METEOR
Mean Pool	0.389	0.362	0.580	0.263
LSTM	0.385	0.347	0.578	0.261
Mean + LSTM	0.388	0.364	0.575	0.259
Temp Att	0.382	0.368	0.580	0.258
Mean + Temp Att	0.385	0.368	0.58	0.26
Ranked Att (ours)	0.387	0.376	0.589	0.264
Mean + Ranked Att (ours)	**0.404**	**0.376**	**0.592**	**0.268**

Comparisons on ranked attention (MSR-VTT)

5b. Analysis & Qualitative Results



Word distribution analysis for generated captions in MSR-VTT testing set

Qualitative results using our STaTS model.

A collage of video frames and captions generated by our ST, TS, STaTS models (above)

Top: Video frames, Below: Spatial Attention visualized on the frame selected by TS attention and the respective word generated.