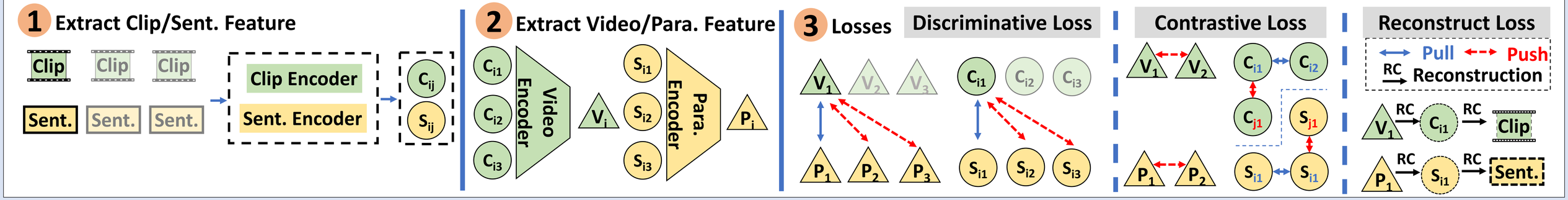
Cross-Modal and Hierarchical Modeling of Video and Text

Bowen Zhang[1*], Hexiang Hu[1*], and Fei Sha[2]

[1]University of Southern California, [2]Netflix

ECCV 2018 European Conference on Computer Vision

Approach



Highlights

- Propose to hierarchically model cross-modal sequential data.
- Preserve correspondence of complex structures across modalities through discriminative losses and contrastive losses.
- State-of-the-art performance on video and paragraph retrieval.
- Systematical study on several tasks involving video and language.

Goal

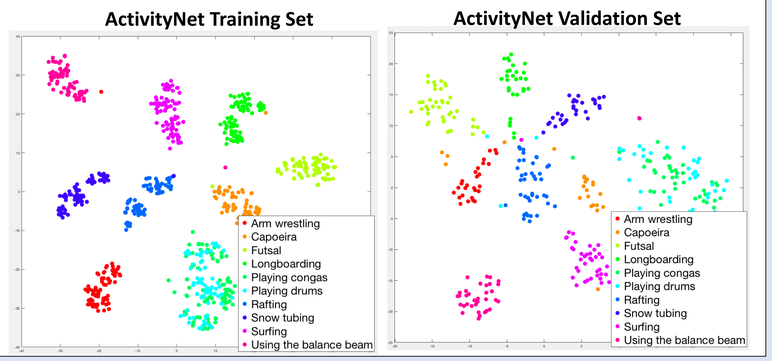
- Learn embeddings for hierarchical sequential data (**video** and **text**) where they have correspondence across multiple modalities.

Tasks & Datasets

- **Tasks:** Video/Text Retrieval, Video Captioning, Zero-shot Action Recognition
- **Datasets:** ActivityNet Dense Caption; ActivityNet V1.3; DiDeMo

Qualitatively Results

- T-SNE visualization of video embedding of **HSE** on ActivityNet V1.3



Experiments & Analysis

- **Video and Text Retrieval:** **With** Ground-truth clip proposal

Table1. Performance on ActivityNet Dense Caption

	Paragraph => Video			Video => Paragraph		
	R@1	R@5	R@50	R@1	R@5	R@50
C3D with Dimension Reduction						
DENSE	14.0	32.0	65.0	18.0	36.0	74.0
FSE	12.6	33.2	77.6	11.5	31.8	77.7
HSE	**32.7**	**63.2**	**90.8**	**32.8**	**63.2**	**91.2**
Inception-V3						
FSE	18.2	44.8	89.1	16.7	43.1	88.4
HSE	**44.4**	**76.7**	**97.1**	**44.2**	**76.7**	**97.0**

Table2. Performance on DiDeMo

	Paragraph => Video			Video => Paragraph		
	R@1	R@5	R@50	R@1	R@5	R@50
Inception-V3						
S2VT	11.9	33.6	76.5	13.2	33.6	76.5
FSE	13.9	36.0	78.9	13.1	33.9	78.0
HSE	**29.7**	**60.3**	**92.4**	**30.1**	**59.2**	**92.1**

Our approach **HSE** outperforms SotA by a large margin.

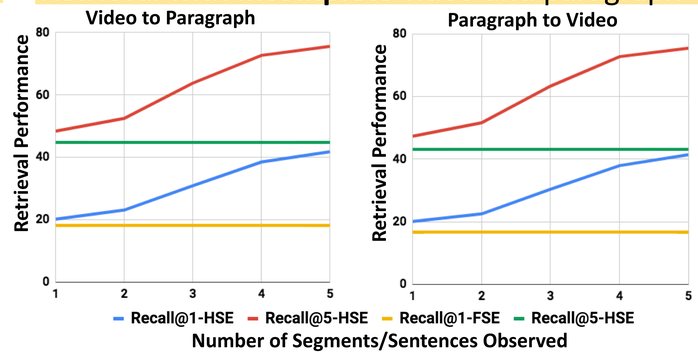
- **Ablations:** **With** heuristic clip proposal

Table3. Performance on ActivityNet Dense Caption w/o clip proposal

Proposal Method		Paragraph => Video		Video => Paragraph	
Inception-V3	#Seg.	R@1	R@5	R@1	R@5
FSE	-	18.2	44.8	16.7	43.1
HSE+GT	-	44.4	76.7	44.2	76.7
HSE + Uniform	**3**	**20.0**	**48.6**	**18.2**	**47.9**
HSE + Uniform	**4**	**20.5**	**49.3**	**18.7**	**48.1**

With a poor uniform proposal, **HSE** can already outperform **FSE** methods.

Retrieval with **incomplete** video and paragraph



- **Video Captioning and Zero-shot Action Recognition:**

Table 4. Results for video captioning on ActivityNet

	B@1	B@2	B@3	Meteor	CiDER
DENSE	26.5	13.5	7.1	9.5	24.6
DVC	19.6	9.9	4.6	10.3	25.2
FSE	17.9	8.2	3.6	8.7	**32.1**
HSE	19.8	9.4	4.3	9.2	**39.8**

Table 5. Results for action recognition on ActivityNet

	Zero-shot Transfer		Train Classifier	
	Top-1	Top-5	Top-1	Top-5
FV-VAE	-	-	78.6	-
TSN	-	-	88.1	-
FSE	48.3	79.4	74.4	94.1
HSE	51.4	83.8	75.3	94.3

- **Check paper for more results and ablations studies!**