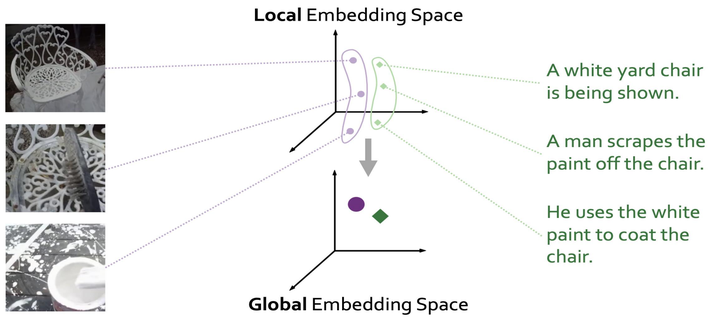
COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning

Mohammadreza Zolfaghari [1], Simon Ging [1*], Hamed Pirsiavash [2], Thomas Brox [1]*
[1]University of Freiburg, [2]University of Maryland Baltimore County
[1]{gings, zolfagha, brox}@cs.uni-freiburg.de, [2]hpirsiav@umbc.edu

Overview

Given a video and a paragraph describing the video, the goal is to learn representations that are semantically aligned in the embedding space.
We propose a hierarchical model that can exploit long-range temporal context both in videos and text when learning the joint cross-modal embedding. We learn a local embedding space and then transform them into a global embedding space.

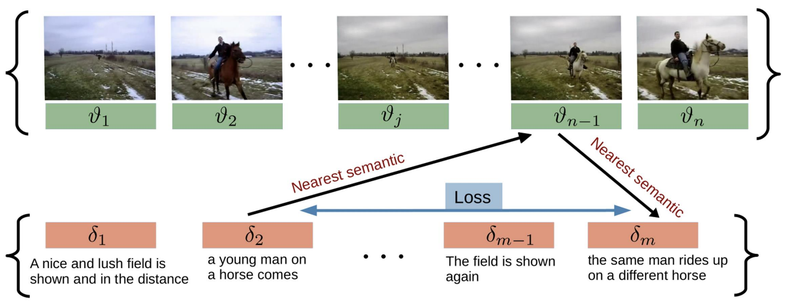


Cooperative hierarchical Transformer (COOT) leverages this hierarchy information with three major components:

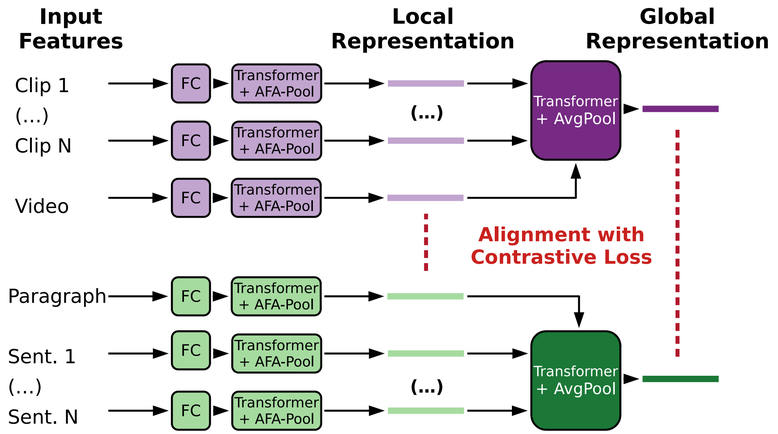
- **Cross-modal Cycle-consistency Loss:** Enforce interaction between modalities and encourage the semantic alignment between them in the learned common space.
- **Intra-level Cooperation:** An attention-aware feature aggregation layer to focus on temporal interactions between low-level entities.
- **Inter-level Cooperation: A** contextual attention module, which enforces the network to highlight semantics relevant to the general context of the video and to suppress the irrelevant semantics.

Cross-modal Cycle-consistency Loss

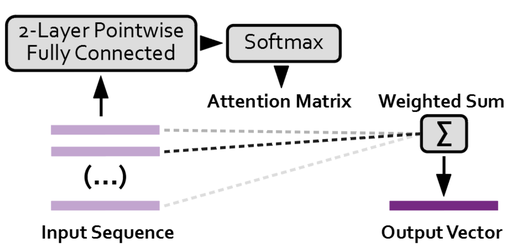
- Use **Temporal Ordering** to enforce cross-model interaction
- **Cycle** between neighbours in clip-sentence space
- Loss: Squared index distance between the start and end point of the cycle.



Cooperative Hierarchical Transformer (COOT)



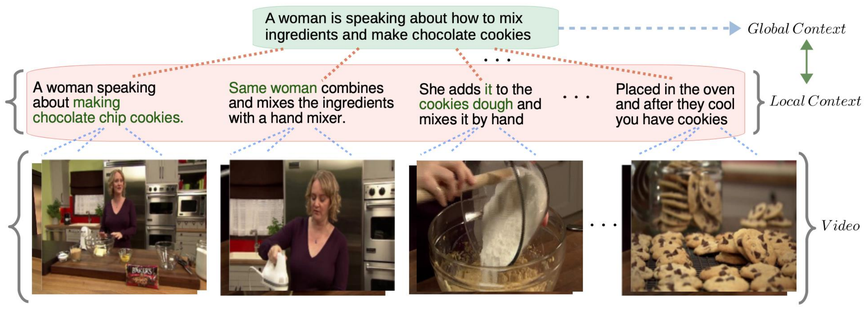
Intra-Level Cooperation (Attention-aware feature aggregation)



This attention-based pooling method helps to consider the relationship between sequence elements to highlight the relevant features.

We achieve this by attending to specific features depending on the local context.

Inter-Level Cooperation (Contextual Transformer)



By modeling the interactions between local and global context, the network learns to highlight semantics relevant to the general context of the video and to suppress the irrelevant ones.

As shown in above figure, without knowing the global context, just from observing the frame in the third clip, there is no information about what type of "dough" is involved. Also the "same woman" in the second clip could not be related to the woman seen in the first clip.

Results: Video-Text Retrieval

Query: A young person goes snowboarding on a mountain. He goes down, jumps and flips and spins.

Results:

ActivityNet-Captions (val split, Paragraph to Video)

Method	Published	R@1	R@5	R@50	MR
CMHSE	ECCV (2018)	44.4±0.5	76.7±0.3	97.1±0.1	2
COOT (Ours)	NeurIPS (2020)	**60.8**±0.6	**86.6**±0.4	**98.6**±0.1	**1**

Youcook2 (val split)

		Par. ⟹ Video		Sent. ⟹ Clip	
Method	Published	R@1	R@5	R@1	R@5
Miech et al.	ICCV (2019)	59.6	86.0	8.2	24.5
ActBert	CVPR (2020)	-	-	9.6	26.7
MIL-NCE	CVPR (2020)	61.9	89.4	15.1	38.0
COOT (Ours)	NeurIPS (2020)	**77.2**±1.0	**95.8**±0.8	**16.7**±0.4	**40.2**±0.3

Results: Video Captioning

Ground Truth: Chop *celery*, *apple*, *red grapes* and *roasted walnuts*. Whisk *mayonnaise*, *lemon juice* and *pepper* and combine with the *fruits* and *nuts*. Place the *salad* on *lettuce*.
MART, ACL (2020): Chop some *red onion and garlic*. Add the *beef* to a bowl and mix. Add *cabbage* and *cabbage* to the bowl.
COOT (Ours): Chop the *celery root* and add them to a bowl. Add *lemon juice*, *olive oil*, *salt* and *pepper* to the bowl and mix well. Toss the *salad*.

ActivityNet-Captions (ae-test split)

Features	Method	TrainSet	B@3	B@4	RougeL	METEOR	CIDEr-D	R@4↓
RGB+Flow	VTransformer	ActivityNet	16.27*	9.31	29.18*	15.54	21.33	7.45
RGB+Flow	TransformerXL	ActivityNet	16.71*	10.25	30.53*	14.91	21.71	8.79
RGB+Flow	MART	ActivityNet	16.43*	9.78	30.63*	15.57	22.16	5.44
COOT video+clip	VTransformer	ActivityNet	16.80	10.47	30.37	15.76	25.90	19.14
COOT video+clip	MART	ActivityNet	**17.43**	**10.85**	**31.45**	**15.99**	**28.19**	6.64

Youcook2 (val split)

Features	Method	TrainSet	B@3	B@4	RougeL	METEOR	CIDEr-D	R@4↓
RGB+Flow	VTransformer	YouCook2	13.08*	7.62	32.18*	15.65	32.26	7.83
RGB+Flow	TransformerXL	YouCook2	11.46*	6.56	30.78*	14.76	26.35	6.30
RGB+Flow	MART	YouCook2	12.83*	8.00	31.97*	15.90	35.74	**4.39**
COOT	VTransformer	H100M$^{\triangle}$+YC2	17.79	11.05	37.51	19.79	55.57	5.69
COOT	MART	H100M$^{\triangle}$+YC2	**17.97**	**11.30**	**37.94**	**19.85**	**57.24**	6.69

Our paper on arXiv:
https://arxiv.org/abs/2011.00597

PyTorch code and models:
https://github.com/gingsi/coot-videotext

References: Check our paper on arXiv for full references. Credit of the bird image to Laurie Boyle - Australia.