End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Presented by Bang, Lorry, Luchao, Xinyu

UNC | COLLEGE OF ARTS AND SCIENCES
Computer Science

Motivation

Vision and language play an important role in the way humans learn to associate visual entities to abstract concepts and vice versa. This has also become the de facto way to successfully train computer vision models.

Limitation: it requires manually annotating large collections of visual data.

Motivation

Limitation: the scale of fully supervised video datasets

Our approach: leveraging narrated videos that are readily available at scale on the web - HowTo100M

Motivation

Limitation: the weak alignment between the video content and the narrative language

Our approach: A bespoke training loss, dubbed MIL-NCE is proposed, enabling the learning to cope with the highly misaligned narration descriptions

Introduction

Train video representations **from scratch** with the novel training scheme MIL-NCE and a simple joint video and text embedding model

The representations obtained are **competitive** with their strongly supervised counterparts on four downstream tasks over eight video datasets.

Prior Work

Task: Learning visual representations from unlabeled videos

Prior: Collect **metadata from social media as supervision**; often in the form of keywords or tags rather than natural language; often platform dependent and rarely publicly available

Our work: Define a supervised proxy task using labels directly generated from videos **(self-supervised) by automatic speech recognition (ASR)**

Prior Work

Prior: Rely on **manually annotated** image / video description datasets, or leverage representations already **pre-trained** on manually labelled datasets (e.g. ImageNet or Kinetics); Do not model any misalignment issue encountered when training

Our Work: **No manually annotated visual data is involved** at any stage of our approach; Address visually **misaligned narrations** from uncurated videos

Overview

Inputs: n pairs of video clips (3.2 seconds each in experiments) and associated narration (16 words max in experiments)

Goal: learn a joint embedding space where the video and text embeddings are similar when the video and text contents are semantically similar

Proposed method: **M**ultiple **I**nstance **L**earning - **N**oise **C**ontrastive **E**stimation (MIL-NCE)

Background Concepts

- Multiple Instance Learning: arrange training data in bags, each bag has a binary label (the goal is to predict unseen bags)

- Noise Contrastive Estimation: loss function that enables comparing positive and negative sample pairs

Method

- Simple joint probabilistic model: $p(x, y; f, g) \propto e^{f(x)^\top g(y)}$

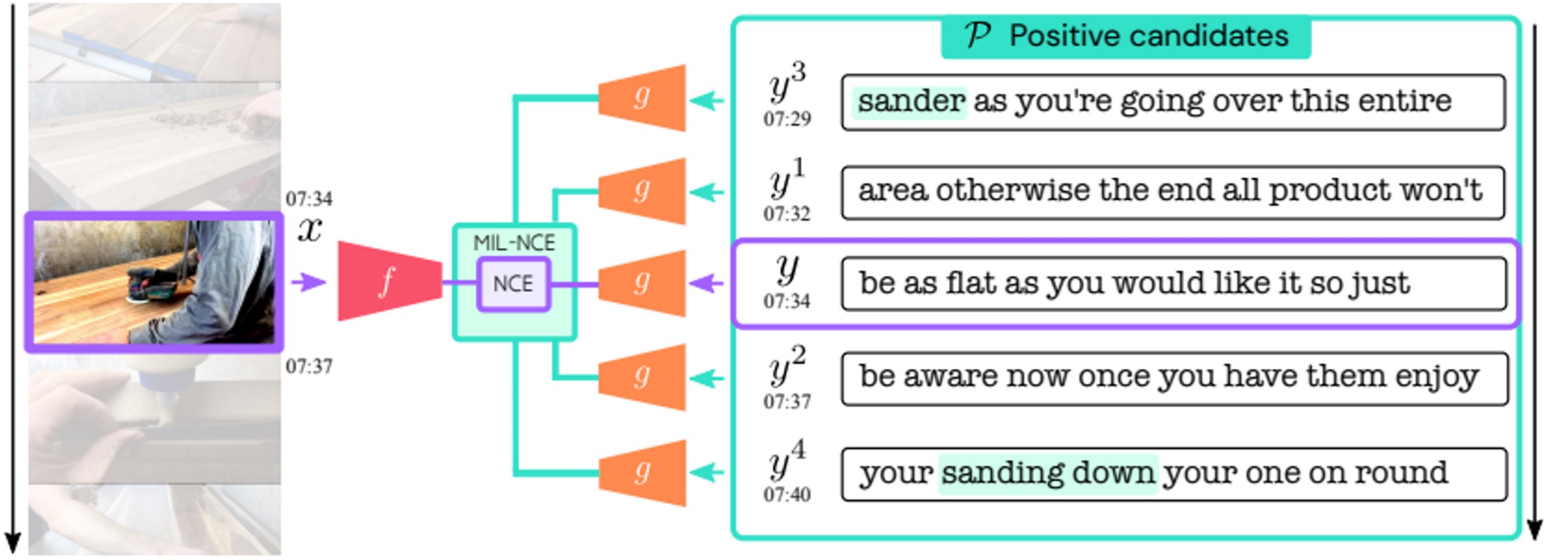
  only considers a single video with a single narration (50% of the HowTo100M dataset are not aligned)

- MIL-NCE: consider the K narrations closest in time as positive candidates to increase the chance that narration correlates to video content

  new joint probabilistic: $p(\cup_k\{(x, y_k)\}) = \sum_k p(x, y_k) \propto \sum_k e^{f(x)^\top g(y_k)}$

- even more generally: $p(\mathcal{P}) \propto \sum_{(x,y)\in\mathcal{P}} e^{f(x)^\top g(y)}$

Method



(a) **Examples of positive candidates**

Method

The MIL-NCE Objective:

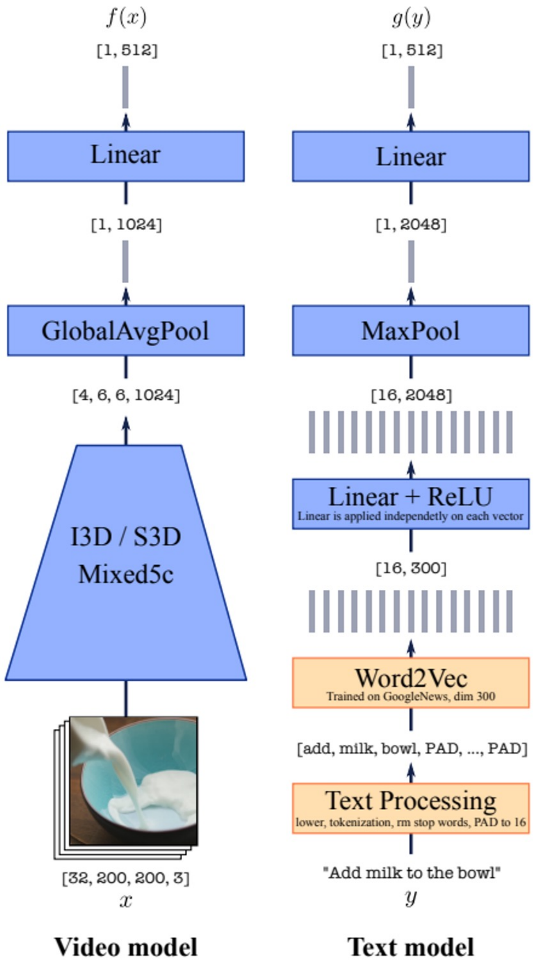
$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum\limits_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

P: positive candidate sets where the nearest narrations in time are selected

N: negative candidate sets that are artificially sampled with $\{(x_i, y_j)\}_{i\neq j}$

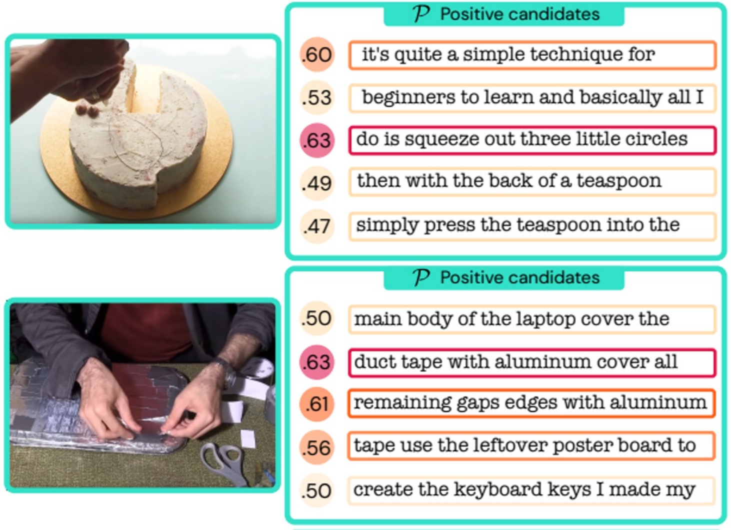
Implementation

- Video Model
  - Input: 3.2 seconds clips (32 frames at 10 fps)
  - I3D/S3D
- Text Model
  - Input: max length of 16
  - Word2Vec + Language Model
- Train on HowTo100M



Downstream Tasks

- Action Recognition: HMDB-51, UCF-101, Kinetics-700
- Text-to-Video retrieval: YouCook2, MSR-VTT
- Action Localization: YouTube-8M Segments
- Action Step Localization: CrossTask
- Action Segmentation: COIN



Ablation Studies

(a) **Training loss**

Loss	YR10	MR10	CTR	HMDB	UCF
Binary-Classif	18.5	23.1	32.6	44.2	68.5
Max margin	16.3	24.1	29.3	**56.2**	76.6
NCE	**29.1**	**27.0**	**35.6**	55.4	**77.5**

(b) **Negatives per positive**

$\|\mathcal{N}\|$	YR10	MR10	CTR	HMDB	UCF
64	26.0	25.5	33.1	56.1	76.0
128	27.1	26.4	33.3	**57.2**	76.2
256	28.7	28.7	**36.5**	56.5	**77.5**
512	**28.8**	**29.0**	35.6	55.4	77.4

(c) **Number of positive candidate pair**

	NCE	MIL-NCE				
$\|\mathcal{P}\| \rightarrow$	1	3	5	9	17	33
YR10	29.1	33.6	**35.0**	33.1	32.4	28.3
MR10	27.0	30.2	**31.8**	30.5	29.2	30.4
CTR	35.6	**37.3**	34.2	31.8	25.0	25.0
HMDB	55.4	**57.8**	56.7	55.7	54.8	51.4
UCF	77.5	79.7	**80.4**	79.5	78.5	77.9

(d) **MIL strategy**

Method	YR10	MR10	CTR	HMDB	UCF
Cat+NCE	31.9	30.8	**35.2**	56.3	78.9
Max+NCE	32.3	31.3	32.2	55.3	79.2
Attn+NCE	32.4	30.2	33.4	55.2	78.4
MIL-NCE	**35.0**	**31.8**	34.2	**56.7**	**80.4**

(e) **Symmetric vs asymmetric negatives**

Negatives	YR10	MR10	CTR	HMDB	UCF
$(x\|y)$	34.4	29.0	33.9	55.1	78.1
$(y\|x)$	19.3	19.4	28.2	**57.1**	79.2
(x, y)	**35.0**	**31.8**	**34.2**	56.7	**80.4**

(f) **Language models**

Text model	YR10	MR10	CTR	HMDB	UCF
LSTM	16.6	15.6	23.8	53.1	80.1
GRU	16.8	16.9	22.2	54.7	**82.8**
Transformer	26.7	26.5	32.7	53.4	78.4
NetVLAD	33.4	29.2	**35.5**	51.8	79.3
Ours	**35.0**	**31.8**	34.2	**56.7**	80.4

Comparison to the state-of-the-art

Method	Dataset	MM	Model	Frozen	HMDB	UCF
OPN [46]	UCF	✗	VGG	✗	23.8	59.6
Shuffle & Learn [54]*	K600	✗	S3D	✗	35.8	68.7
Wang *et al.* [78]	K400	Flow	C3D	✗	33.4	61.2
CMC [74]	UCF	Flow	CaffeNet	✗	26.7	59.1
Geometry [25]	FC	Flow	FlowNet	✗	23.3	55.1
Fernando *et al.* [24]	UCF	✗	AlexNet	✗	32.5	60.3
ClipOrder [86]	UCF	✗	R(2+1)D	✗	30.9	72.4
3DRotNet [37]*	K600	✗	S3D	✗	40.0	75.3
DPC [30]	K400	✗	3D-R34	✗	35.7	75.7
3D ST-puzzle [40]	K400	✗	3D-R18	✗	33.7	65.8
CBT [71]	K600	✗	S3D	✓	29.5	54.0
CBT [71]	K600	✗	S3D	✗	44.6	79.5
AVTS [43]	K600	Audio	I3D	✗	53.0	83.7
AVTS [43]	Audioset	Audio	MC3	✗	**61.6**	89.0
Ours	HTM	Text	I3D	✓	**54.8**	**83.4**
				✗	59.2	**89.1**
			S3D	✓	**53.1**	**82.7**
				✗	61.0	**91.3**
Fully-supervised SOTA [85]			S3D	✗	75.9	96.8

Method	Labeled dataset used	R@1↑	R@5↑	R@10↑	MedR↓
Random	None	0.03	0.15	0.3	1675
HGLMM FV CCA [42]	ImNet + K400 + YouCook2	4.6	14.3	21.6	75
Miech *et al.* [52]	ImNet + K400	6.1	17.3	24.8	46
Miech *et al.* [52]	ImNet + K400 + YouCook2	8.2	24.5	35.3	24
Ours (I3D)	**None**	**11.4**	**30.6**	**42.0**	**16**
Ours (S3D)	**None**	**15.1**	**38.0**	**51.2**	**10**

(a) **YouCook2**

Method	Labeled dataset used	R@1↑	R@5↑	R@10↑	MedR↓
Random	None	0.01	0.05	0.1	500
Miech *et al.* [52]	ImNet + K400	7.5	21.2	29.6	38
Ours (I3D)	**None**	**9.4**	**22.2**	**30.0**	**35**
Ours (S3D)	**None**	**9.9**	**24.0**	**32.4**	**29.5**

(b) **MSRVTT**