Weakly Supervised Representation Learning for Audio-Visual Events

Sanjeel Parekh[1,2] Slim Essid[2] Alexey Ozerov[1] Ngoc Duong[1]
Patrick Pérez[1] Gaël Richard[2]

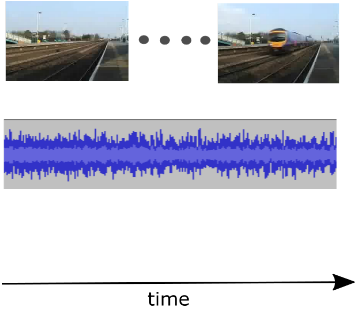
[1]Technicolor, France

[2]Télécom ParisTech, Université Paris-Saclay, France

September 2018

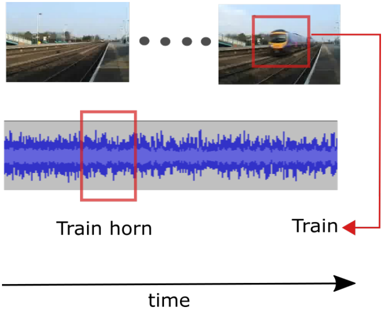
Goal

Given a video of an audio-visual event (AV)



Goal

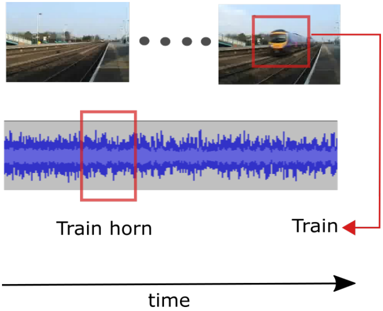
Given a video of an audio-visual event (AV)



- Which AV event has occurred?
- Where is the visual object/context that distinguishes the event?
- When does the sound event occur?

Goal

Given a video of an audio-visual event (AV)



- Which AV event has occurred?
- Where is the visual object/context that distinguishes the event?
- When does the sound event occur?

> – Events in the two modalities may be asynchronous
> – Only video-level labels available, without any timing information

Related Works

Object Localization and Classification

- Embedding multiple instance learning (MIL) strategies in CNN architectures
  (Oquab et al., 2015; Zhou et al., 2016; Kolesnikov and Lampert, 2016)
- MIL over extracted region proposals
  (Bilen and Vedaldi, 2016; Kantorov et al., 2016; Gkioxari et al., 2015)

Related Works

Object Localization and Classification

- Embedding multiple instance learning (MIL) strategies in CNN architectures
  (Oquab et al., 2015; Zhou et al., 2016; Kolesnikov and Lampert, 2016)
- MIL over extracted region proposals
  (Bilen and Vedaldi, 2016; Kantorov et al., 2016; Gkioxari et al., 2015)

Audio Event Detection

- Recent progress accelerated by introduction of large datasets e.g. AudioSet by Google (Gemmeke et al., 2017)
- Success of deep MIL and attention-based methods for audio
  (Xu et al., 2017; Kumar et al., 2017)

Related Works

Object Localization and Classification

- Embedding multiple instance learning (MIL) strategies in CNN architectures
  (Oquab et al., 2015; Zhou et al., 2016; Kolesnikov and Lampert, 2016)
- MIL over extracted region proposals
  (Bilen and Vedaldi, 2016; Kantorov et al., 2016; Gkioxari et al., 2015)

Audio Event Detection

- Recent progress accelerated by introduction of large datasets e.g. AudioSet by Google (Gemmeke et al., 2017)
- Success of deep MIL and attention-based methods for audio
  (Xu et al., 2017; Kumar et al., 2017)

Multimodal Deep Learning

- Two-stream architectures (nonlinear feature-space transformation methods)
  (Becker and Hinton, 1992; Aytar et al., 2016; Aytar et al., 2017; Arandjelović and Zisserman, 2017; Andrew et al., 2013; Ngiam et al., 2011),
- Cross-modal architectures
  (Yuhas et al., 1989; Owens et al., 2016a; Owens et al., 2016b)

Proposed Approach

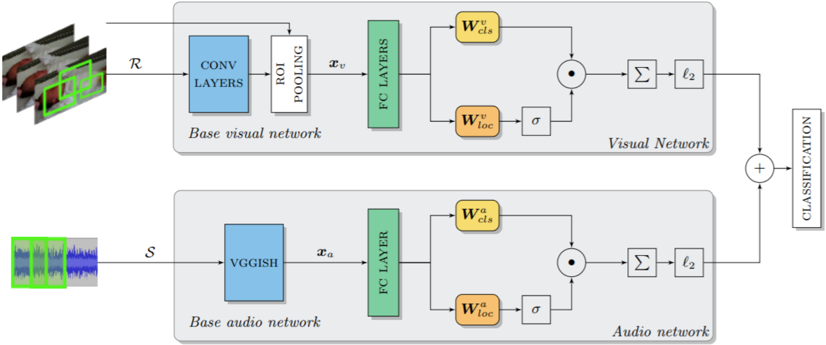
Key Idea: Propose and Learn

- Consider each video to be a bag of class-agnostic audio and visual proposals
- Extract features and transform them to score each according to their relevance for a particular class

Proposed Approach

Key Idea: Propose and Learn

- Consider each video to be a bag of class-agnostic audio and visual proposals
- Extract features and transform them to score each according to their relevance for a particular class



Proposed Approach

Given a set of N training videos and labels, $\{V^{(n)}, y^{(n)}\}$, both modules are jointly trained using multilabel hinge loss

$$L(w) = \frac{1}{CN} \sum_{n=1}^{N} \sum_{c=1}^{C} \max\Big(0, 1 - y_c^{(n)} \phi_c(V^{(n)}; w)\Big). \tag{1}$$

Experiments

- We use the DCASE smart cars challenge data, which is a subset of AudioSet
  - Multi-label dataset with 51,172 training samples, 488 validation and 1103 testing samples
  - 17 classes spread over vehicle sounds (e.g. bus, car, truck) and warning sounds (e.g. car alarm, civil defense siren)

Experiments

- We use the DCASE smart cars challenge data, which is a subset of AudioSet
  - Multi-label dataset with 51,172 training samples, 488 validation and 1103 testing samples
  - 17 classes spread over vehicle sounds (e.g. bus, car, truck) and warning sounds (e.g. car alarm, civil defense siren)

- Baselines
  1. Visual-only network (Bilen and Vedaldi, 2016)
  2. Audio-only network
  3. AV One-Stream Architecture using a *log-sum-exponential* operator
  4. Attention-based CVSSP system (Xu et al., 2017): DCASE smart cars challenge winner ; uses no external data

Audio Event Classification Results

System	F1	Precision	Recall
Proposed AV Two Stream	**64.2**	59.7	**69.4**
TS Audio-Only	57.3	53.2	62.0
TS Video-Only	47.3	48.5	46.1
TS Video-Only WSDDN-Type (Bilen and Vedaldi, 2016)	48.8	47.6	50.1
AV One Stream	55.3	50.4	61.2
CVSSP - Fusion system (Xu et al., 2017)	55.6	**61.4**	50.8
CVSSP - Gated-CRNN-logMel (Xu et al., 2017)	54.2	58.9	50.2

- Significantly advance state-of-the-art for event classification

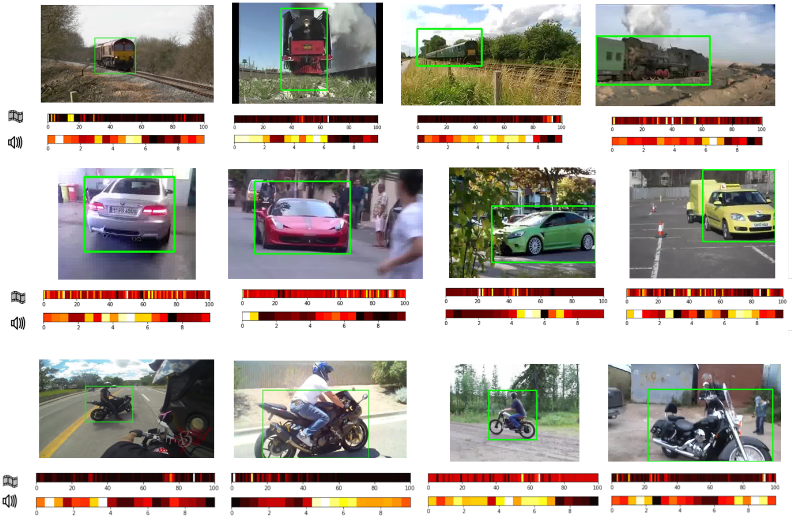
Audio Event Classification Results

System	F1	Precision	Recall
Proposed AV Two Stream	**64.2**	59.7	**69.4**
TS Audio-Only	57.3	53.2	62.0
TS Video-Only	47.3	48.5	46.1
TS Video-Only WSDDN-Type (Bilen and Vedaldi, 2016)	48.8	47.6	50.1
AV One Stream	55.3	50.4	61.2
CVSSP - Fusion system (Xu et al., 2017)	55.6	**61.4**	50.8
CVSSP - Gated-CRNN-logMel (Xu et al., 2017)	54.2	58.9	50.2

System	Vehicle Sounds								Warning Sounds								
	bik	bus	car	car-pby	mbik	skt	trn	trk	air-hrn	amb	car-alm	civ-def	f-eng	pol-car	rv-bps	scrm	trn-hrn
Proposed AV TS	**75.7**	**54.9**	**75.0**	**34.6**	**76.2**	**78.6**	82.0	**61.5**	40.0	**64.7**	53.9	80.4	64.4	49.2	36.6	81.1	47.1
TS Audio-Only	42.1	38.8	69.8	29.6	68.9	64.9	78.5	44.0	40.4	58.2	53.0	79.6	61.0	51.4	42.9	72.1	46.9
TS Video-Only	72.5	52.0	61.2	15.0	54.1	64.2	73.3	49.7	12.0	33.9	13.5	68.6	46.5	19.8	21.8	44.1	32.1
AV OS	68.2	53.6	74.1	25.6	67.1	74.4	**82.8**	52.8	28.0	54.7	20.6	76.6	60.4	56.3	18.8	49.4	36.2
CVSSP - FS	40.5	39.7	72.9	27.1	63.5	74.5	79.2	52.3	**63.7**	35.6	**72.9**	**86.4**	**65.7**	**63.8**	**60.3**	**91.2**	**73.6**

- Significantly advance state-of-the-art for event classification
- Audio-visual complementarity

Localization Results



- Video of localization examples

  We can effectively deal with unsynchronized events

Summary

- Proposal-based formulation allows for symmetric treatment of both modalities through MIL
- Established effectiveness of multimodal approach: AV complementarity and tackling asynchronous events
- Ongoing work on designing task-specific proposals for problems such as audio source separation

Thank you!