Sparse in Space and Time: Audio-visual Synchronisation with Trainable Selectors

Vladimir Iashin [1] Weidi Xie [2,3] Esa Rahtu [1] Andrew Zisserman [3]

[1]Tampere University [2]Shanghai Jiao Tong University [3]University of Oxford

Goal

Audio-visual synchronisation of videos with **sparse** cues

Challenges

- Sync signal is rare → longer input sequences
- Absence of a dataset with sparse sync cues
- Hidden temporal artefacts in data → model learns a shortcut

Contributions

1. Novel multi-modal transformer architecture, **SparseSync**
   - Scales linearly with respect to input length
   - Predicts the offset size
2. Study of the video codec compression artefacts
   - MPEG-4 Part 2 (`mpeg4`) and AAC leak temporal artefacts
   - Recommendation: avoid `mpeg4` and use H.264, 16kHz AAC is ok
3. Video dataset with sparse sync signals, **VGGSound-Sparse**
   - We also suggest benchmarking future models on "uncropped" LRS3

Datasets

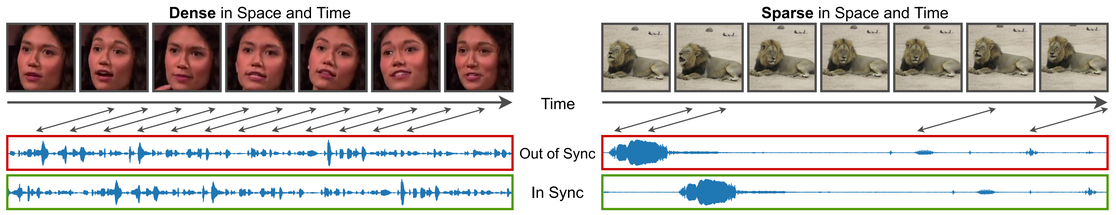
VGGSound-Sparse

- New video dataset with sparse sync signals
- 12 classes from VGGSound (6.5k videos, 10 seconds)
- e.g. *dog barking*, *chopping wood*, *striking bowling*
- "**Sparse** in time and **sparse** in space"

LRS3-H.264 (uncropped scene)

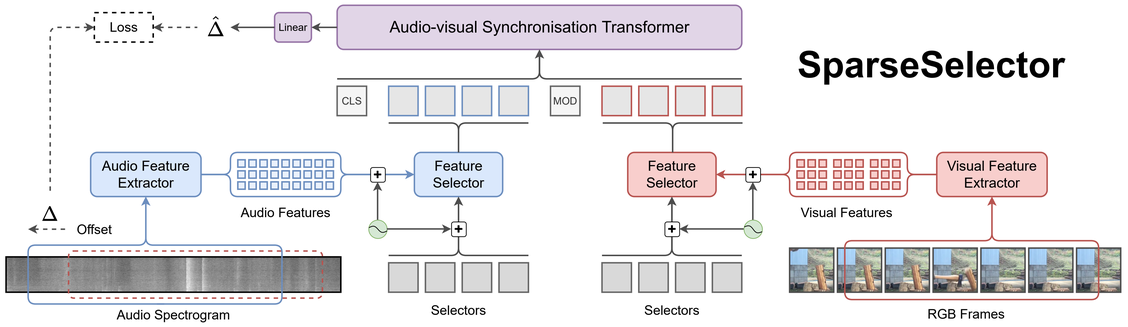
- 58k clips from 4.8k TED presentations
- As LRS3 (Afouras *et al.*, 2018) but uncropped and in H.264
- "**Sparse** in space but **dense** in time"

Dense vs. Sparse Sync Signals



- *Easy:* talking heads interviews (left)
- *Difficult:* open-domain classes with sparse sync signal (right)

Synchronising Videos with Sparse Signals



SparseSelector

Overview

1. Features are extracted from spectrogram and RGB frames
2. Trainable selectors *query* sync cues from audio and visual features via cross-attention
3. Audio and visual tokens are concatenated
4. Sync transformer predicts the temporal offset for synchronisation

Training

- Offset classification: (−2.0, −1.8, ..., 0.0, ..., +2.0) — 21 classes
- Offsets are random and made on-the-fly
- 5-second clips from 10-second videos
- Pre-train on dense signals (LRS3-H.264) → fine-tune on sparse signals (VGGSound-Sparse)

Results

	LRS3 (no crop) Accuracy	*VGGS-Sparse* Accuracy
$AVST_{dec}$	83.1	29.3
Ours	96.9	44.3

$AVST_{dec}$ is an adaptation of (Chen *et al.*, 2021)

Improving Performance

Increase Input Length

Length (sec.)	3	4	5	6	7
Accuracy	36.8	43.0	44.3	45.6	46.5

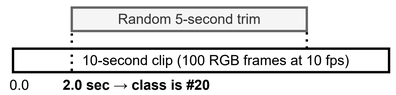
Pre-training on non-sparse data-classes

Pre-training	Accuracy
LRS3 (no crop) + VGGSound-Sparse	44.3
LRS3 (no crop) + VGGSound (full)	51.2

Evaluated on test-set of VGGSound-Sparse

Temporal Artefacts

Train a model to predict the start of the crop



- the model should not train but it does
- both audio and visual streams are affected