Learning ambient sounds embeddings

Nicolas Turpault, Romain Serizel, Emmanuel Vincent

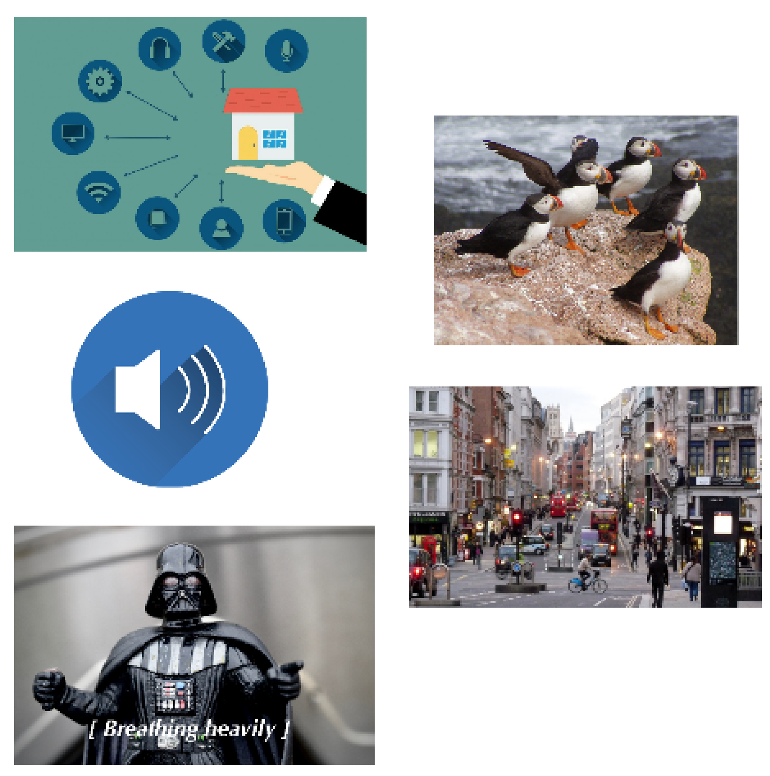
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

Overview

- Learning a **representation (embeddings)** of ambient sounds.
- Defining the problem (weak labels, multilabels, unbalanced data).
- Work on semi-supervised learning and identify problem of weakly labeled data.

Ambient sounds, why ?

- **Domestic sounds**: home assisted living, smart home, security
- Urban sounds: urbanisation
- Animal sounds: migratory phenomena
- Audio captioning
- Sound library (similar sounds)



Problem definition

Audio: Time-Frequency representation.

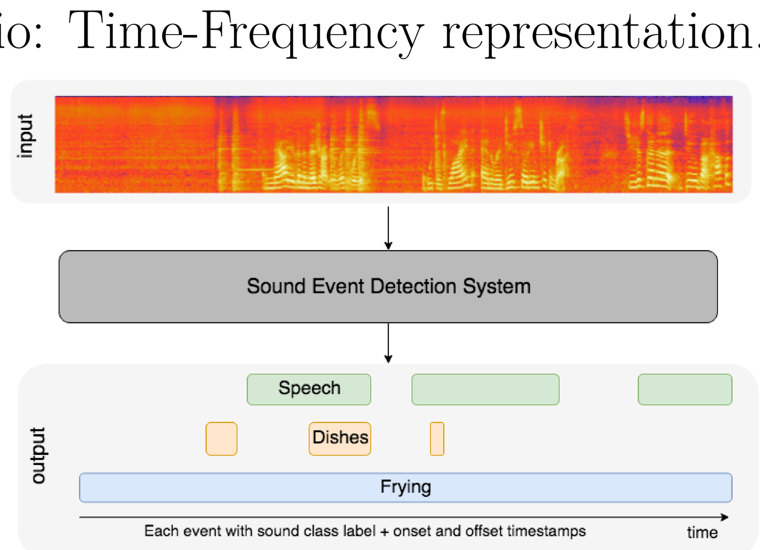


Figure 1: Sound event detection

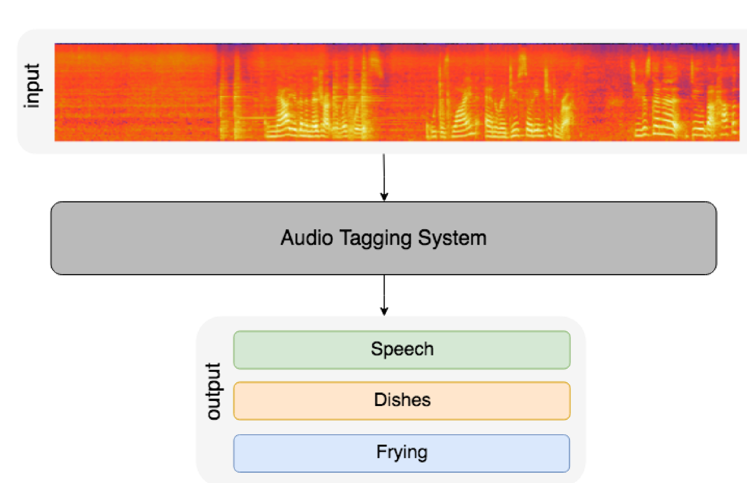


Figure 2: Audio tagging

- No temporal information (weak labels)
- Representation + Classification learned

- Temporal information (strong labels)
- Representation + Classification learned

Representation: Time consuming, **common** for multiple applications.
Classifier: Problem dependent.

Data

- 10 event classes (unbalanced).
- Multilabel.
- Overlapping sounds.
- **Weakly labeled data.**
- Synthetic data (domain mismatch).
- **Semi supervised learning.**
- Big variation of length of events.

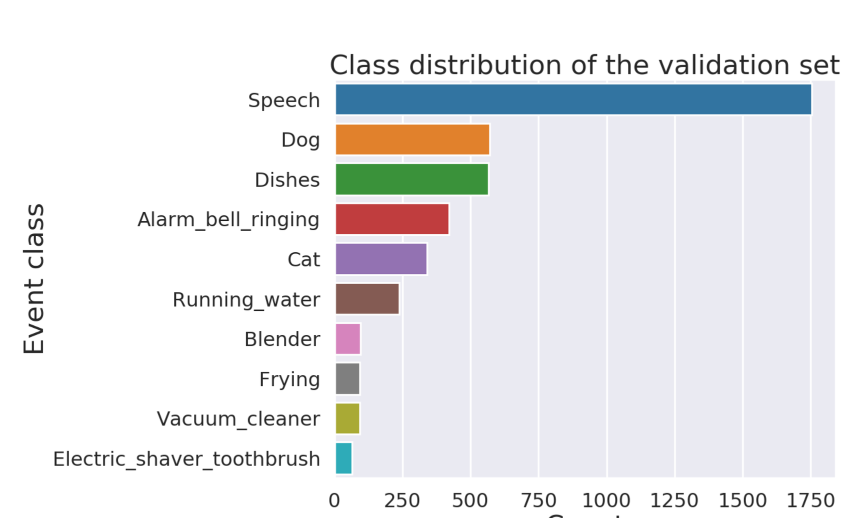


Figure 4: Unbalanced classes

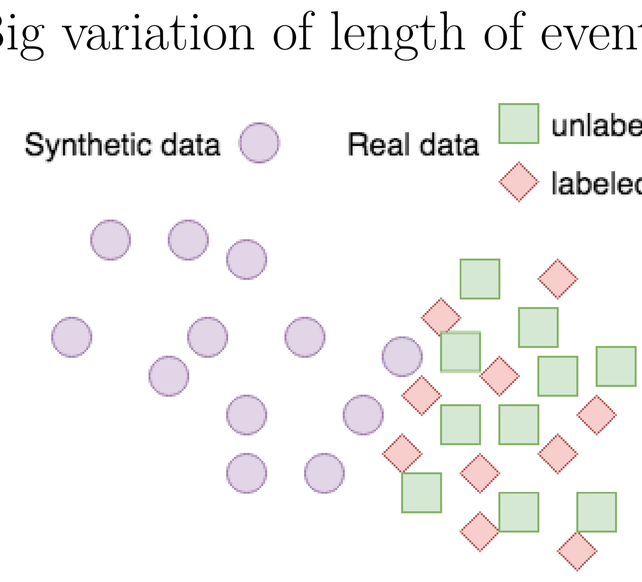


Figure 3: Domain mismatch

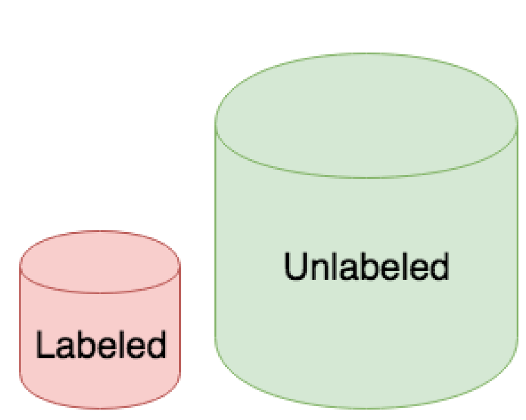


Figure 5: Semi supervised data

Learning embeddings

Problem studied: **semi supervised learning**: lot of unlabeled data available, **weakly labeled data**: low resources annotations.
Method used: **triplet network** (sampling method).
Triplet: Anchor (reference), positive (similar label with the anchor, or augmented version of the anchor), negative (label different with the anchor).

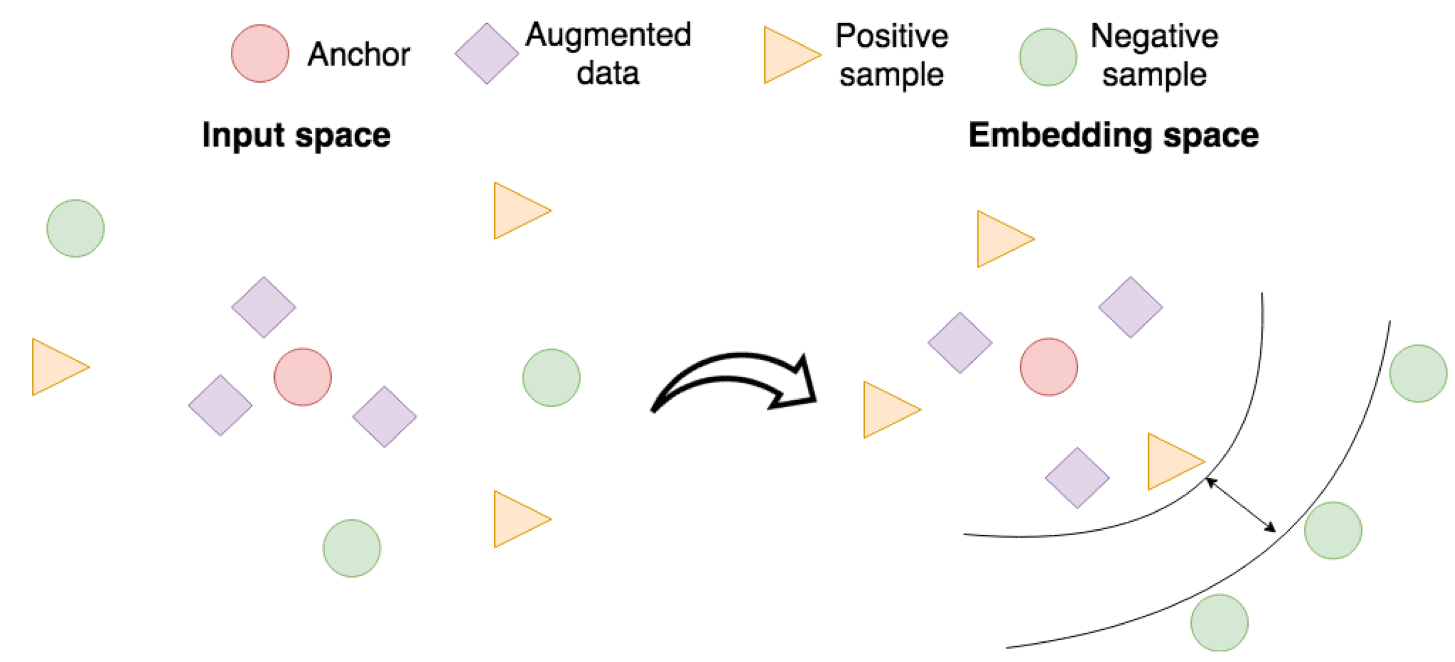


Figure 6: Triplet learning objective.

Results

Embeddings evaluated on a classification task (audio tagging).
Semi supervised problem
Fully supervised training: 53.6% mean macro F-score.

Nb unlabeled	*7,890*	*15,780*	*19,725*	*23,670*
Nb labeled	*23,670*	*15,780*	*11,835*	*7,890*
Positive augmented (%)	55.2±0.7	54.4±0.7	47.3±6.2	17.4±26.8

Table 1: Macro F1-score (%), on the evaluation set. Varying number of labeled (L) and unlabeled (U) triplets. 95% confidence score over 3 launches.

Weakly labeled data (synthetic data only)

Method	Training Time (s)	Testing time (s) 0.2	1.0	10.0
Triplets	0.2	42.5±1.0	38.2±3.6	11.7±3.2
	1.0	41.7±7.0	44.8±10.9	18.3±7.3
	10.0	9.1±3.2	10.2±2.0	2.8±0.7

Table 2: F-measure results on the WAA2 dataset (in %)

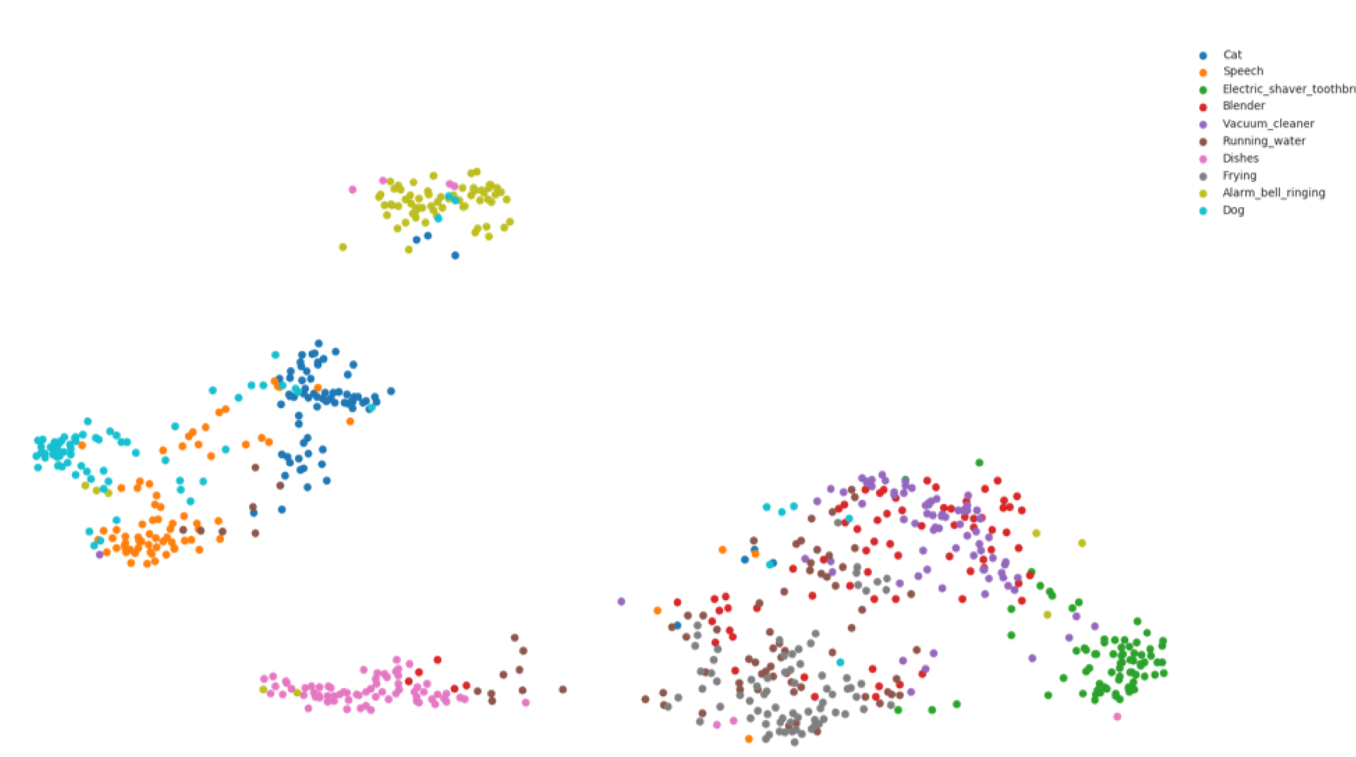


Figure 7: T-SNE (non-linear 2D) representation of our 10 classes using 1 sec per point.

Conclusion and future work

- Benefit of semi-supervised training.
- Explained the problem of weakly labeled data (to be overcome).
- There is semantic information in the embeddings.
- Length of events matters (will be studied).
- Segmentation is important (will be studied).

This work was made with the support of the French National Research Agency, in the framework of the project LEAUDS "Learning to understand audio scenes" (ANR-18-CE23-0020) and the French region Grand-Est. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations.