Improved Representation Learning for Acoustic Event Classification using Tree-Structured Ontology

A. Zharmagambetov[1], Q. Tang[2], C.-C. Kao[2], Q. Zhang[2], M. Sun[2], V. Rozgic[2], J. Droppo[2], C. Wang[2]

[1]University of California, Merced [2]Amazon Alexa

Poster Number: 4605

1 Abstract

Acoustic events have a hierarchical structure analogous to a tree (or a directed acyclic graph). In this work, we propose a structure-aware semi-supervised learning framework for acoustic event classification (AEC). Our hypothesis is that the audio label structure contains useful information that is not available in audios and plain tags. We show that by organizing audio representations with a human-curated tree ontology, we can improve the quality of the learned audio representations for downstream AEC tasks. We use consistency training to use large amounts of unlabeled data for structured representation manifold learning. Experimental results indicate that our framework learns high quality representations which enable us to achieve comparable performance in discriminative tasks as fully supervised baselines. Moreover, our framework can better handle audios with unseen tags by confidently assigning a super-category tag to the audio.

2 Introduction

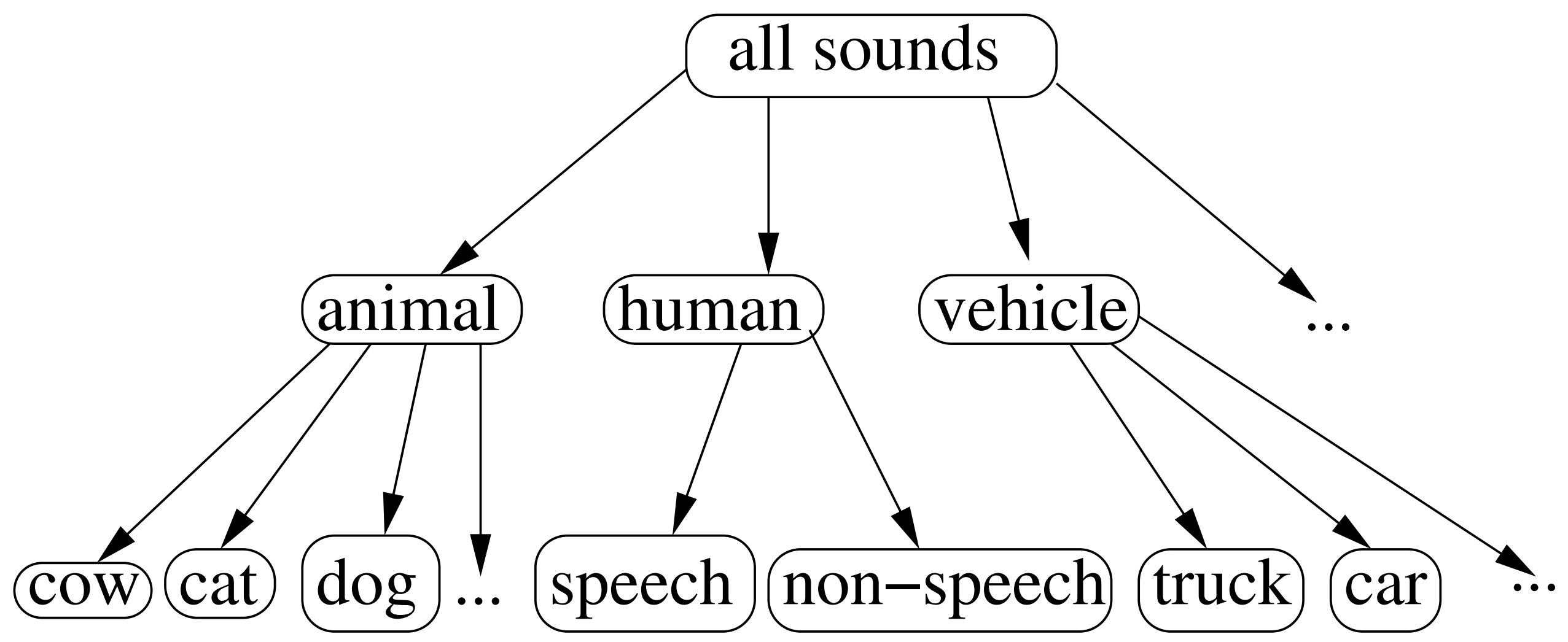
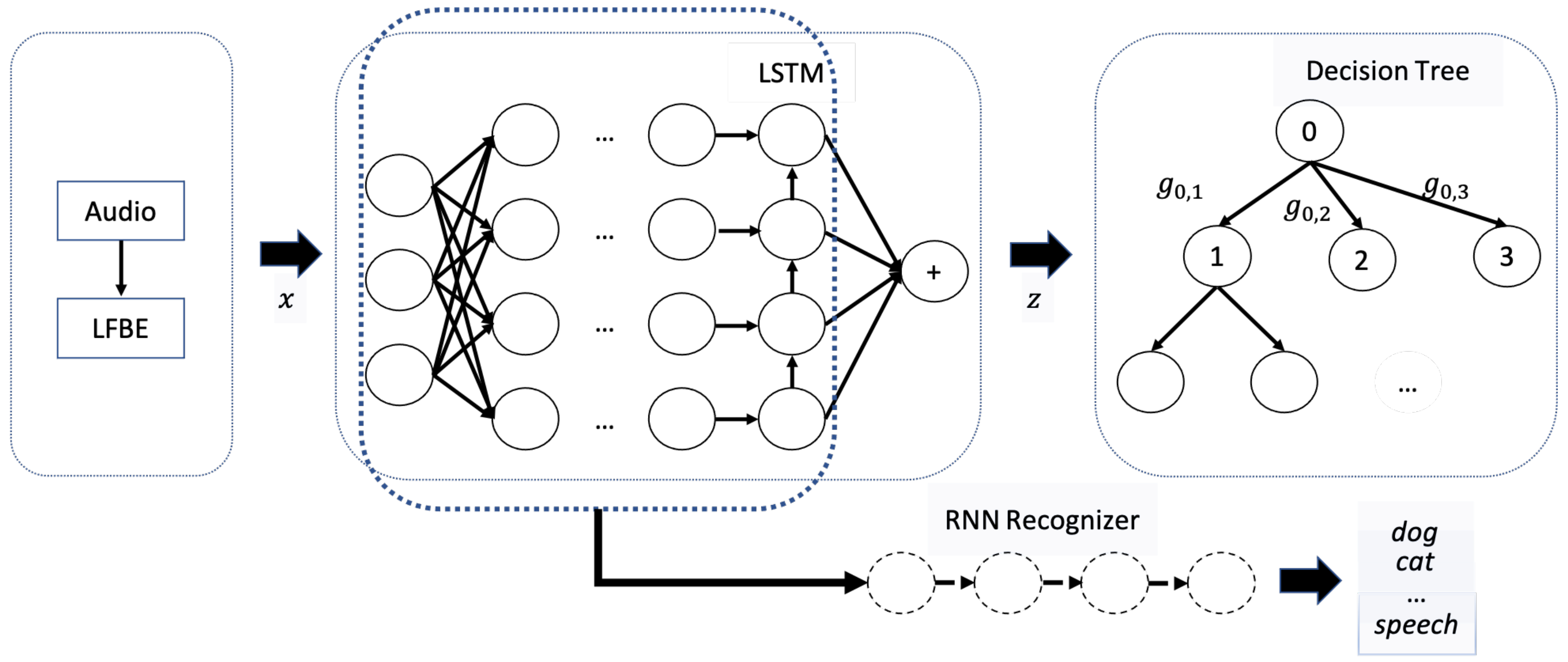


Figure: Example of the tree-structured ontology of various audio events. Hierarchical structure allows to group similar audio events under the same “super-category”.

- We consider representation neural networks pretrained on a large amount of data (mostly unlabeled).
- Extracted representations can be used by a wide range of downstream tasks, e.g. acoustic event classification.
- We capitalize on the idea of using a tree-structured ontology (see figure above) to guide the training of the representation network.
- Why Tree?
  - Availability of the human curated ontology;
  - Well suited for annotation purposes;
  - Can be applied to detect/categorize novel sounds;
  - Encouraging the audio representation manifold to align with tree ontology organizes representations in more reasonable fashion;
  - etc.

3 Tree-structured ontology for representation learning



- Representation network (encoder) takes log mel-filterbank energy (LFBE) $\mathbf{x}$ as input and produces an embedding vector $\mathbf{z} = f(\mathbf{x}; \Theta)$. Decision tree operates on embedding space and has its own learnable parameters at each node. The encoder + decision tree is trained jointly; for downstream tasks (like RNN recognizer), the encoder is frozen.
- For each tree node i we define a gating function $g_i(\cdot)$ with trainable parameters $\mathbf{W}_i, \mathbf{b}_i$:
$$g_i(\mathbf{z}; \mathbf{W}_i, \mathbf{b}_i) = \sigma(\mathbf{W}_i^T \mathbf{z} + \mathbf{b}_i) \qquad (1)$$
  The above softmax function outputs probability distribution over the children nodes of i. Then, the probability of reaching a certain leaf:
$$P(y|\mathbf{z}) = g_{\mathbf{c}_0,\mathbf{c}_1}(\mathbf{z}) \cdot g_{\mathbf{c}_1,\mathbf{c}_2}(\mathbf{z}) \dots g_{\mathbf{c}_{l-1},\mathbf{c}_l}(\mathbf{z}) \qquad (2)$$
  where $g_{\mathbf{c}_{i-1},\mathbf{c}_i}$ indicates the probability of transition from node $\mathbf{c}_{i-1}$ to its child $\mathbf{c}_i$ along root-to-leaf path.
- Training: given a training set (X, Y), we minimize the negative log-likelihood loss shown below:
$$L_s(\mathbf{W}, \Theta) = -\mathop{\mathbb{E}}_{\mathbf{x},y} \frac{\sum_{j=1}^{|y|} \log P(y_j|f(\mathbf{x}))}{|y|} \qquad (3)$$
- How to leverage unlabeled data?
  We use consistency training. Assuming we are given a portion of data indexed in set $\{\mathbf{x}_m\}_{m=1}^{M}$, the unsupervised consistency loss is defined as:
$$L_c(\mathbf{W}, \Theta) = \mathop{\mathbb{E}}_{\mathbf{x}} \left\{ \mathop{\mathbb{E}}_{\mathbf{c} \in \mathbf{C}} \{\mathcal{D}(g_c(\mathbf{z}), g_c(\hat{\mathbf{z}}))\} \right\} \qquad (4)$$
  where $\mathbf{z}$ and $\hat{\mathbf{z}}$ are the representations of $\mathbf{x}$ and $\hat{\mathbf{x}}$ (perturbed duplicate of $\mathbf{x}$), respectively. To obtain $\hat{\mathbf{x}}$, we use common augmentation methods, such as SpecAugment which applies transformations on the LFBE surface feature.
- The final loss:
$$L(\mathbf{W}, \Theta) = L_s(\mathbf{W}, \Theta) + \lambda L_c(\mathbf{W}, \Theta) \qquad (5)$$

4 Experiments

- Comparison of different methods on a downstream classification task (on a subset of AudioSet) given fixed representations. F1 score is given as the average of 5 runs.

Method	Label fraction 1%	 5%	 10%
	Test F1 macro		
supervised	———	0.612	———
supervised (with aug.)	———	0.638	———
LFBE	0.201	0.442	0.508
SimCLR+APC	0.244	0.491	0.551
SimCLR+APC → Fine-tune	0.329	0.540	0.587
SDT	0.288	0.531	0.580
SDT + consistency (SDTC)	0.410	0.557	0.600
SDTC + APC (SDTCA)	**0.417**	**0.561**	**0.609**

- We can use the tree itself (along with representation network) to measure the average accuracy by levels where “Level 1” checks if a leaf node is correctly classified, parent and grandparent for “Level 2” and “Level 3”. Consider as an example audio clips of “dog barking”. We collect all audio clips which belong to this event. We measure how many of them are correctly classified as “dog” (Level 1), as “domestic animal”(Level 2), etc.

	Seen		Unseen	
Avg. acc.	SDT	Baseline	SDT	Baseline
Level 1	0.46	0.35	-	-
Level 2	0.61	0.44	0.59	0.41
Level 3	0.79	0.62	0.78	0.57

Conclusion

- We leverage a tree-structured ontology of audio events for representation learning for acoustic event classification.
- To achieve this goal, we propose a parametric tree model which can be jointly trained with a representation encoder.
- We apply a semi-supervised learning scheme based on consistency training that can be used to handle labeled data scarcity issue.
- To the best of our knowledge, this is the first attempt to use consistency training for tree-based models in the AEC domain.
- Experimental results suggest that:
  - SDT-based semi-superivsed learning can improve AEC performance by conveying the structural information hidden in the label ontology to learned audio embeddings.
  - The proposed approach allows to more confidently classify unseen/novel events (at the level of super-categories).