Specialized Embedding Approximation for Edge Intelligence: A Case Study in Urban Sound Classification

Sangeeta Srivastava[1], Dhrubojyoti Roy[1], Mark Cartwright[2], Juan Pablo Bello[2], and Anish Arora[1]
[1]The Ohio State University
[2]Music and Audio Research Laboratory, New York University

The Ohio State University

SONYC Sounds of New York City

NYU

Poster Number: 3967

1. The Problem

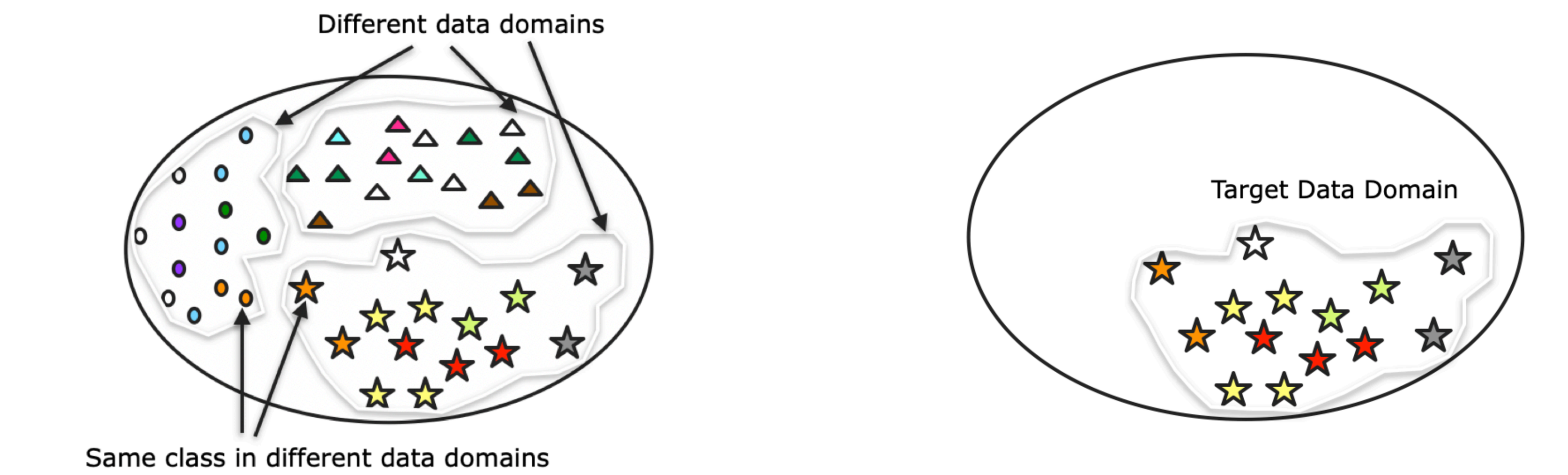
- Audio embedding models for edge devices
- Limited compute, memory, and storage in edge devices
  - 1MB of RAM and 2MB of Flash in ARM Cortex-M7 Microcontroller
- Large embedding models
  - L3-Net audio 18 MB in size and requires 12 MB of activation memory

2. Limitations

- Traditional knowledge distillation uses teacher data to train student net
  - Teacher's data to achieve both cross- and intra-domain generalizability
- Sub-optimal compression
  - Student net's training more complex than necessary
- Necessitates availability of teacher's train data

3. Our Solution: Specialized Embedding Approximation

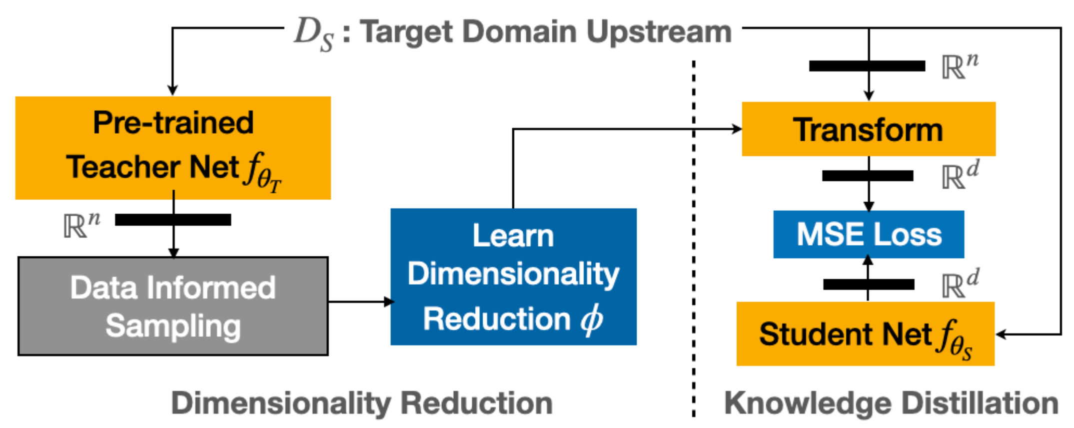
- Audio embedding models for edge devices **for a target domain**
- Train the student to approximate only the portion of the teacher's embedding manifold relevant to the target domain



Teacher's Embedding Space: Cross- and Intra-domain generalizable

Student's Embedding Space: Intra-domain generalizable

- Block diagram of the SEA pipeline:



4. Case Study: Urban Sound Classification

- Sounds of New York City (SONYC) aims at continuous monitoring, analysing, and mitigating urban noise pollution
- Upstream data
  - Unlabeled recordings from 15 sensors placed in New York
  - Audio + Sound Pressure Level (SPL)
- Downstream data: SONYC-UST
  - Multi-label dataset of 3068 annotated 10-second audio recordings

5. SEA Student Nets

- Reduced input representation
  - 8 kHz sampling, 64 mel filters instead of L3's 48 kHz, 256 mels
- Reduced architecture
- Trained on SONYC upstream with SEA

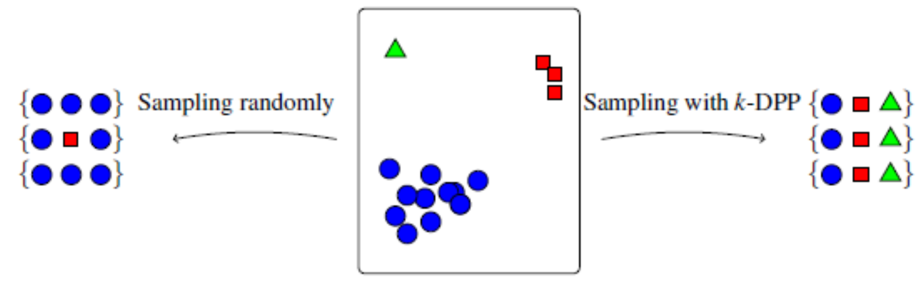
6. Evaluation on SONYC-UST

Model	Emb. Dim.	Model Size (MB)	Act. Mem. (MB)	Micro-AUPRC
L3-Audio	512	18.80	12.79	0.810
Student 0	512	18.80	0.82	0.823
Student 1	256	4.70	0.41	0.793
Student 2	128	2.34	0.41	0.797
Student 3	64	1.64	0.41	0.784

- SONYC SEA L3: 8-bit quantized Student 2
  - **Flash: 0.585 MB** and **RAM: 0.1025 MB**
- Train Efficiency
  - 10x lesser train data
  - converges 5x (10x) faster with a learning rate of 10^{-5} (10^{-4})
- Compromise out-of-domain performance

7. Dimensionality Reduction with Informed Sampling

- Reduce memory overhead during training of dimensionality reduction
- Sampling types
  - Random
  - Only Relevance: SPL informed
  - Only Diversity | Diversity + Relevance: Determinantal Point Process
- *Relevance*: More informative data points
  - Higher relative loudness → potential noise source
- *Diversity*: Diverse set to capture most of the structure information in desired domain



Sampling Type	Micro-AUPRC
Diversity + Relevance	0.783
Random	0.782
Only Diversity	0.781
Only Relevance	0.779

- SONYC SEA students used PCA reduction with Diversity + Relevance

8. edgel3

- **pip install edgel3**

```
import edgel3
import soundfile as sf

audio, sr = sf.read('/path/to/file.wav')

# Get embedding out of SEA Student 2 (UST data domain)
emb, ts = edgel3.get_embedding(audio, sr, model_type='sea', emb_dim=128)

# Get embedding out of 95.45% sparse L3
emb, ts = edgel3.get_embedding(audio, sr, model_type='sparse', sparsity=95.45)
```

https://github.com/ksangeeta2429/embedding-approx

SCAN ME