Methods to Overcome Data Scarcity

Bruno Padovese
Problem Statement

- Deep Learning has been shown to provide excellent results
  - Text (translation, negotiations...)
  - Voice identification and Recognition
  - Video (object identification, lip reading...)
  - Image classification and generation

- However, a major bottleneck exists
  - Deep Learning requires large amounts of data!

Source: feedthedatamonster.com
Datasets in Modern Deep Learning

CIFAR 10

- Very high quality benchmarks datasets
Data quality is also critical in Machine Learning applications

- In reality, most datasets will be very far away from ideal

Collecting data is a hard and time-consuming task. High quality data, doubly so

In the realm of bioacoustic it gets even harder

- Remote, inhospitable locations
- Manpower heavy, even with the help of Passive acoustic monitoring
- Costly
Bioacoustics also presents an additional bottleneck: annotating large amounts of raw data

- Multiple sensors from PAM systems can produce an insurmountable amount of data to manually label
- Only a fraction of the data is actually useful
- Many Datasets could only been 1% annotated

How can we mitigate this problem? Collect more data? Annotate more data?

Recently, several approaches have been proposed to diversify our datasets
Data Augmentation
Data Augmentation

- Data Augmentation are techniques that can be employed to increase the size and diversity of our datasets
  - Ways to artificially generate data while still being realistic

- Can be used to increase performance of models to generalize
  - Introduces new data to the training set

- Methods range from trivial to complex, both presenting good results
Data Augmentation - Shifting time domain
Data Augmentation - Spec Augment
When dealing with Audio, augmentation techniques can be applied in the spectrogram domain or directly to the waveform or both. Shifting and SpecAugment are examples of spectrograms augmentation. Techniques applied to the waveform are also widely used.
Data Augmentation - Mixup

Humpback Whale

Northern Right Whale
Data Augmentation - Mixup
● Changing the spectrogram generation parameters can also yield different representations of the same signal
  ○ Different spectral resolutions to be able to distinguish different spectral features

● Augmentation has become a default pre-processing step in many fields
  ○ DCASE - Detection and Classification of Acoustic Scenes and Events
  ○ SpecAugment and Mixup being used in most works
Data Augmentation

- Simple methods that significantly increases performance of deep learning models

- Data Augmentation should not and will not replace real data!
  - Efforts towards collecting and labelling more data should continue
  - Data Augmentation tries to fill gaps present in current datasets
  - Augmented data is used alongside real data in training procedures
  - Methods are only as good as the data they try to match
Few-shot Learning

**Meta-learning**
Learn a learning strategy to adjust well to a new few-shot learning task

**Metric learning**
Learn a semantic embedding space using a distance loss function

**Data augmentation**
Synthesize more data
GANs (Generative Adversarial Networks) and WaveNet
Generative Models
Generative Models
Generative Models

Input random variable (drawn from a simple distribution, for example uniform).

The generative network transforms the simple random variable into a more complex one.

Output random variable (should follow the targeted distribution, after training the generative network).

The output of the generative network once reshaped.

Source: https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29
Generative Models - WaveNets
Non dilated Causal Convolutions
Generative Models - WaveNets
Generative Models - WaveNets
Will these types of method work well in bioacoustics with little modification?
○ Can it be used to generate a whale upcall for example?
GANs

Forward transform of the initial random variables to generate data

Backpropagation of the matching error to train the network

Input random variables (drawn from a uniform).

Generative network to be trained.

The generated distribution is compared to the true distribution and the “matching error” is backpropagated to train the network.

Source: https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29
GANs

- True Dataset
- True Samples
- Discriminator
- Generated Samples
- Backpropagate Error
- Random noise
- Generator
- True Samples
- Fake
- True

Backpropagate Error
Data Augmentation
Questions

- How good do these generators need to be?
- How much data do you need to achieve good generations?
- How much augmentation do you need? What type?
How does it relate to Meridian?
Workflow visions

MERIDIAN’s library of pre-trained models

- model 1
- model 2
- model 3

Performance

Scenario

User’s scenario

- right whales
- right whales + ships
- right whales + humpback whales
- right whales + seismic noise
Workflow visions

Transfer learning

model 2 + Data + generated data = custom model

Performance

Scenario

User’s scenario

right whales + seismic noise
Thank you