Our task :
21 target species

1500 species DCNN :

Models:

Until 2016, the best system used template matching of small sound elements combined with a large number of acoustic features and a random forest ensemble learning method for classification [4],[5].

Features to be carful about : sample rate, bit depth, number of channels,

1st step :-

In a first step, two data sets are formed with homogeneous file properties. For the first set all files are resampled to 44.1 kHz followed by normalization to a

maximum signal amplitude of -3 dB. For a second set all files are first high pass filtered at a frequency of 2 kHz (Q = 0.707), resampled to 22050 Hz, mixed to mono

and finally also normalized to -3 dB. The training set is augmented by additionally extracting 381 audio files using the time coded annotation of species in the metadata of the newly provided soundscape validation set. All files belonging to the training set are further processed to create additional data sets with different content described more detailed below:
BirdsOnly
NoiseOnly
AtmosOnly
LowQuality

Via segmentation each training file is is separated as signal and noise parts
Segmentation is done in frequency domain applying image processing

methods like **median clipping** [8] and further morphological operations on the spectrogram image to extract individual sound events. And use them for **template matching**

Xeno-Canto

Some files are convertyed to mp3 encoding with lame parameter V7
IMagenet vs yolo

Types of loss functions:

The basic data loading pipeline can

be summarized as follows:
extract audio chunk from file with a duration of ca. 5 seconds
apply short-time Fourier transform
normalize and convert power spectrogram to decibel units (dB) via logarithm
convert linear spectrogram to mel spectrogram
remove low and high frequencies
resize spectrogram to fit input dimension of the network
convert grayscale image to RGB image

entropy is used as loss function considering only foreground species as ground truth targets. Stochastic gradient descent is used as optimizer with Nesterov momentum of

0.9, weight decay of 1e-4 and an initial learning rate of 0.01. Learning rate is decreased during training by ca. 10-1

in 3 to 4 steps until 0.0001 whenever performance

on the validation set stopped improving.

Paper 2:
Dataset: cornell bird challenge 2020 with 100/264 different species (https://xeno-canto.org.)

Species that are common for our task:
(not done)

Note:CNN it does not take into consideration the underlying
temporal dependence characteristics of the species calls.

Stratified 5fold
Librosa for mel spectrogram

Visualizing using PCA on softmax distribution

Models used:
ML models that can be considered : hidden markov, random forest, SVM
Deep Learning VGG16, resnets

Best model : (CNN) on a slice of the spectrogram, and a Recurrent Neural Network (RNN) for the **temporal component to combine across time-points (acc=67%)**